Chapter 2

An Approach to the Identification and Phylogenetic Analysis of Emerging and Hemorrhagic Fever Viruses

Francisco J. Díaz, Luis E. Paternina, and Juan David Rodas

Abstract

An important aspect of virological surveillance is the identification of the detected viruses. Broad surveillance, that typically employs deep sequencing of collected tissue samples, provides the investigator with many sequence files constructed from overlapping stretches of DNA sequences. Directed surveillance for viruses of a specific taxonomic group provides the investigator with sequence files from cDNA amplified using specific primers to conserved viral regions. Here we will describe general approaches to identify hemorrhagic viral agents through phylogenetic analysis of cDNA sequences obtained during surveillance activities.

Key words Hemorrhagic viruses, Phylogenetics, Databases, Alignment, Genetic trees, Bioinformatics

1 General Introduction

Viral hemorrhagic fevers (VHF) refer to a group of diseases caused by agents from different RNA viral families. The term "viral hemorrhagic fever" is used to describe the multisystem syndrome characterized by impairment of the vascular system, sometimes accompanied by hemorrhage (bleeding). Although the bleeding by itself is rarely the cause of death, and some types of hemorrhagic fever viruses can cause relatively mild illnesses, some of these viruses cause severe, life-threatening disease [1].

Viral hemorrhagic fevers share several features: (1) they all are produced by enveloped RNA viruses; (2) they are all zoonotic, involving transmission by insects, ticks, rodents, bats, or other wild or domestic reservoirs; (3) they are geographically restricted to the areas where their hosts live; (4) humans are usually incidental hosts that are sporadically infected, but in some cases, they can also transmit these viruses to other humans; and (5) outbreaks cannot be easily predicted, and there are very few antiviral treatments and

Maria S. Salvato (ed.), *Hemorrhagic Fever Viruses: Methods and Protocols*, Methods in Molecular Biology, vol. 1604,
DOI 10.1007/978-1-4939-6981-4_2,

vaccines available [1]. So far, most of the agents associated with HF have been found within the following four viral families:

Filoviridae: Ebola and Marburg virus diseases

Arenaviridae: Lassa fever, Lujo, Guanarito, Machupo, Junín, Sabiá, and Chapare viruses

Bunyavirales: Rift Valley fever (mosquito-borne), Crimean-Congo hemorrhagic fever, and hantaviruses

Flaviviridae: Dengue, yellow fever, Omsk hemorrhagic fever, Kyasanur Forest disease, and Alkhurma viruses [2]

Here we will describe freely available programs to align your DNA sequence obtained from surveillance activities with reference sequences deposited in public databases like GenBank; we will describe how to analyze your sequence files in order to get phylogenetic trees for viral agent identification with brief notes about the utility of this work in epidemiological studies (e.g., mapping vector population spread and epidemic start sites).

2 Materials

Besides the sequences, the only other "materials" required for phylogenetic sequence analyses are the appropriate computer hardware and software: a sequence database and one or more phylogenetic packages.

1. Sequence databases. The major sequence database is GenBank, maintained at the National Center for Biotechnology Information (NCBI), Bethesda, Maryland, available at http://www.ncbi.nlm.nih.gov/genbank/ [3]. The European (EMBL) and Japanese (DDBJ) bioinformatic databases are equivalent to GenBank since these three organizations exchange data on a daily basis. Several manuals and tutorials are available at the GenBank web site to get started and to understand the resources of the site. Another useful database for research in hemorrhagic fever viruses is the Virus Pathogen Resource (ViPR) available at http://www.viprbrc.org/. This is a more curated database focused on virus families; among them are *Flaviviridae*, *Togaviridae*, *Arenaviridae*, former *Bunyaviridae*, *Filoviridae*, and others. Besides serving as a repository of sequence data, the site provides several analytical tools for sequence alignment, similarity searches, phylogenetic reconstruction, sequence variation, and more [4].
2. Phylogenetic software. Many computer programs have been developed to perform phylogenetic and related sequence analyses and have been made available. A comprehensive list of them with comments about their uses and links to their web sites is maintained by Professor Joseph Felsenstein at the University of Washington [5]. Most of these are specialized

programs that perform specific analyses, making it necessary to combine them with other programs; so they will not be described here. We recommend MEGA (Molecular Evolutionary Genetic Analysis) software for beginners, a multipurpose, user-friendly, and free software that allows one to perform the full process of bioinformatic and evolutionary analysis of nucleic acid and protein sequences in a single platform. The program includes sufficient help, tutorials, and examples to allow the user to develop familiarity. MEGA is updated frequently; here version 6.0, described in reference [6], will be used in its more basic form. A detailed guide to its use is provided in the book *Phylogenetic Trees Made Easy* [7].

3 Method

The protocol that follows assumes that you have amplified viral sequences, either directly from a patient, host, or environmental sample or from a virus isolate.

1. **Obtain one or several genomic sequences** of your samples or isolates either by direct sequencing a PCR product or from a cloned sequence (*see* **Note 1**). Assemble the products of individual sequencing reactions into a single "contig," and edit conflicting bases or segments. Save your sequence in FASTA format in a plain text editor. Alternatively, sequences obtained by deep (next-generation) sequencing could be used.
2. **Search for homologous sequences** in GenBank. Go to the BLAST page of NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi), and select *nucleotide blast*. Copy and paste your sequence in the Enter Query Sequence window. In Choose Search Set/Database, choose Others "Nucleotide collection (nr/nt)". In Program Selection, select Optimize for "Somewhat similar sequences (blastn)". Check the "Show results in a new window" and click on the BLAST button. The search could take 1 min or more. Look in the Graphic Summary to get a quick idea of the coverage and identity of the homologous sequences found or "hits." Scroll down to the Description section. It is a table with the 100 closest sequences to your query. The first of these hits gives you the identity of the virus in most of the cases. The columns at the right give some metrics showing how well each hit matches to the query sequence. The "Query cover" and "Ident" columns give a percent value of the coverage and identity. The "E-value" is the probability of having this match just by chance. Select the sequences you wish to include in the analysis by checking them at the left. How many and which sequences to select depend on the scope of the analysis, either limited to strains of the same virus or different

species within a genus or a family. As a rule of thumb, avoid selecting sequences with very similar names or with identical E-values because they are probably redundant sequences. Also avoid selecting sequences with a low coverage (say, less than 60–70%). Go back to the top of the table and click on Download and then on FASTA (aligned sequences) to get the dataset of sequences in a text file. Open the dataset with a text editor, include query sequence and (optional) other less closely related sequences to be used as "outgroups" in the analysis. Edit sequence names to a short and meaningful sentence like "Lujo virus South Africa 2008" and save it with .fas extension, say "newseq.fas" (*see* **Note 2**).

3. **Align the sequences**. Open MEGA and note the menu options. Follow this path Align → Edit/Build Alignment → Create New DNA alignment → DNA. The "Alignment Explorer" window will open. Here click the "Open" icon to browse, and open your dataset file. Move your mouse over the toolbar buttons above to know what their functions are. Align the DNA sequences using either Clustal W or MUSCLE. In the window that appears, accept (OK/Compute) the alignment default parameters, unless you have a good reason to modify them. Alignment may take some minutes depending on the size and complexity of your dataset. Visually check the quality of the alignment. If this is satisfactory, go to the Data menu and select "Save session". The program will assign the ".mas" extension (e.g., Newseq.mas). Go back to the Data menu and select Export Alignment → MEGA format. Save this as suggested (e.g., Newseq.meg), and minimize the Alignment Explorer to return to the main menu. Next, use the Data function in the toolbar to open the .meg file you just created. Now you can use the "TA" button to open the "Sequence Explorer" or go directly to the next step.
4. **Select a substitution model**. From MEGA main menu, select Models → "Find best DNA/protein model (ML)". In the incoming window, accept the default Analysis Preferences and compute. These steps may take some minutes depending on your dataset and your computer. It yields a table showing how different substitution models fit to your dataset sorted by the Bayesian information criteria (BIC). The first (upper) model is recommended for the next step (*see* **Note 3**).
5. **Perform the phylogenetic analysis**. With the *.meg file still open, go to the Phylogeny function in Mega main menu, and select Construct/test Neighbor-joining tree. Other methods could also be used, but we recommend trying neighbor-joining first to get a quick phylogenetic tree that could be improved later. In the opening Analysis Preferences window, select the

following options. Test of Phylogeny: Bootstrap method. No. of Bootstrap replications: 1000. Substitution type: nucleotide. Model/method: the model selected in **step 4**. Rates among sites: If the model selected in **step 4** included a +G (e.g., TN93+G), then choose Gamma distributed, and type the value in the (+G) column of the table in the next option (Gamma Parameter); otherwise (e.g., TN93 or TN93+I), select Uniform rates in Rates among Sites. Pattern among Lineages: Same (Homogeneous). Gaps/Missing Data Treatment: Partial deletion. Site Coverage Cutoff (%): 95. Finish with √ Compute. The analysis could take a few seconds or several minutes depending on the dataset (*see* **Note 4**).

6. **Edit and interpret the phylogenetic tree**. When the analysis is finished, a phylogenetic tree will appear in a new window. Move the mouse over the icons in the left and upper bars to explore the editing options available. Look for the out-group sequence(s) in the tree, and click on the branch that leads to it (them); go to the "Place Root on Branch" icon on the left bar, and click on it to place the root properly. Observe the tree carefully. Terminal branches that are too short or too long could indicate redundant sequences or misaligned sequences, respectively. Decide which sequences should be removed and which others are missing, according to your knowledge of the species or genus. Go back to the dataset obtained in **step 2** and modify it properly. When ready, repeat **steps 3–6** until you are satisfied with the tree (*see* **Note 5**).
7. **Consider post-phylogenetic analyses**. Molecular epidemiological study of viruses uses several analytical tools including phylogenetic, phylogeographic, and phylodynamic analyses. The aforementioned stepwise protocol for genetic analysis is the basis of virus evolutionary genetics. Phylogenetic analysis of viral sequences proceeds as outlined in Fig. 1, and each of the steps is described in the protocol with some important notes about it. One of the most prolific uses of such genetic analyses lies in the field of phylogeography. The phylogeographic analysis focuses on the coupled study of the distribution/dispersion process of organisms and their genetic variation. There are several classic examples of phylogeography of hemorrhagic viruses that have provided useful information about their evolution and dispersion patterns [8, 9].

   This kind of work has enabled the mapping of origin points for outbreaks [10] and the dispersion patterns of hemorrhagic fever viruses in their reservoirs [11], and more recently, in conjunction with experimental methods, virologists are trying to make predictions about the future fitness and molecular evolution of some emerging/reemerging viruses [12].

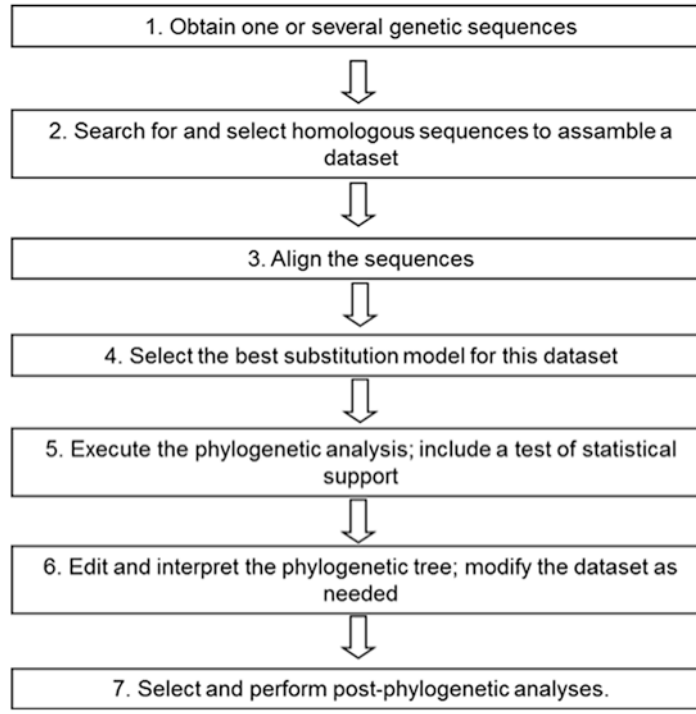

Fig. 1 Steps in a phylogenetic study

4 Notes

1. Both second-generation automated sequencing and third-generation (deep) sequencing technologies are useful for molecular epidemiology purposes. The former can be accomplished by direct sequencing of RT-PCR products or by sequencing recombinant clones of retro-transcribed genomic sequence; this latter variant, however, is not recommended since it could randomly select for sequences that do not represent the majority of the viral population; these sequences could also exhibit additional mutations introduced during the amplification process. On the other hand, direct sequencing yields the consensus sequence of the viral population in the isolate or in the sample. It is worth noting, however, that primers used in diagnostic PCR are not always satisfactory for molecular epidemiology work, since they are often designed to amplify short, well-conserved genomic segments with few variable sites on them. These sequences are good enough for identifying viral species and often for subtyping/genotyping too. However, most powerful phylogenetic and phylodynamic analyses require large alignments of sequences with many variable sites to produce robust, well-supported statistical inferences. The longer

the sequenced segment, the better. Accordingly, sequences of 500 nucleotides or more are sufficient for most analyses, but longer segments are needed when several isolates of an epidemic cluster or from a single endemic site are being studied. Therefore, additional primer pairs amplifying long or contiguous segments are usually required.

Third-generation sequencing technologies are more promising since they allow one to obtain longer, frequently complete, genomic sequences of viral isolates. They also help to identify viral sequences of non-previously identified viruses from clinical or from complex environmental samples [13]. Technical and financial issues still preclude the use of these technologies on a routine basis in most places, but these obstacles could be overcome soon due to the rapid development of these technologies.

2. The first analysis performed on a genomic sequence usually consists of searching for identical or similar sequences in a large bioinformatic database like GenBank, maintained at the National Center for Biotechnology Information (NCBI), Bethesda, Maryland, available at http://www.ncbi.nlm.nih.gov/nuccore/ [3]. BLAST (Basic Local Alignment Search Tool) is the resource available at GenBank for such searches. The "blastn" algorithm, which compares your own (query) nucleotide sequence with all nucleic acid sequences in the database, in sense and antisense directions, is the basic method for searching, but many other algorithms like megablast, blastp, blastx, and others are available for more refined searches in nucleotide and protein databases. Several manuals and tutorials are available at http://www.ncbi.nlm.nih.gov/genbank/ to get started and gain understanding about the resources of the site. BLAST outputs include a table with the description, accession number, percent identity, probability (E-value) of a false-positive hit, and other measures of the most similar sequences. Graphic summaries, alignments, taxonomy reports, several downloading formats, as well as links to the original publication and other resources are also available.
3. When you are performing the substitution model selection, please have in mind that the substitution model option in MEGA only explores the 24 most common substitution patterns, while specialized software for this purpose such as jModelTest has approximately 88 models with other models in development. Because MEGA only works with 24 substitution patterns, just substitution models listed in MEGA are allowed. Model selection: MEGA software typically orders the substitution models according to the highest BIC; however, it also provides AIC and log likelihood for the same substitution model. You need to choose the best model for your data according to your criteria (BIC, AIC, AICc, log likelihood).

4. MEGA offers several kinds of analysis that range from similarity analysis based on dendrograms such as UPGMA, minimum evolution, and neighbor joining up to analysis of characters such as maximum parsimony and maximum likelihood. The tree reconstruction method chosen by the scientist will determine the type of tree that you can obtain, and concordance among them is expected; however, because of the different principles in which they lie on, it is possible to obtain different genetic trees (tree topology, branch length, branch support). Further information about the different methods is provided with examples in *Molecular Evolution and Phylogenetics* [14] and *Phylogenetic Trees Made Easy* [7].

   When you are running your phylogenetic analysis, the most common setting for bootstrapping procedure is 1.000. The bootstrapping is a nonparametric procedure used for assessment of confidence of genetic clustering. In this case, it is used as an estimator for branch support. The result of bootstrap will be a number between zero to a hundred, and the higher the bootstrap (closest to 100), the better supported is your cluster/clade. However, it is important to mention that only the highest values of bootstrap are considered to be good evidence for cluster/clade support and values below 70 are considered non-conclusive support for branches [15].

5. Once you have obtained your tree, you can save this output in *.mts format that allows the saving of not only the phylogenetic tree but also the setup information of the whole analysis (just like a Log file). This information of the analysis contains the substitution model, the options for treatment of gaps, the support method, the number of sites used in the analysis, and the sites used (first + second + third or otherwise).

References

1. CDC (2013) Viral hemorrhagic fevers. CDC Fact Sheet 1–3. doi: 10.1111/j.1749-6632.2009.05106.x
2. Knust B (2016) Chapter 3: viral hemorrhagic fevers. In: Barrett GW (ed) Centers for disease control and prevention: CDC health information for international travel. Oxford University Press, New York, pp 3–6
3. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. Nucleic Acids Res 41:36–42. doi:10.1093/nar/gks1195
4. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, Liu M, Kumar S, Zaremba S, Gu Z, Zhou L, Larson CN, Dietrich J, Klem EB, Scheuermann RH (2012) ViPR: an open bioinformatics database and analysis resource for virology research. Nucleic Acids Res 40:593–598. doi:10.1093/nar/gkr859
5. Felsestein J (1995) University of Washington. http://evolution.genetics.washington.edu/phylip/software.html. Accessed 4 Sept 2016
6. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729. doi:10.1093/molbev/mst197
7. Hall BG (2011) Phylogenetic trees made easy, 4th edn. Sinauer Associates Inc., Sunderland
8. Beck A, Guzman H, Li L, Ellis B, Tesh RB, Barrett AD (2013) Phylogeographic reconstruction of African yellow fever virus isolates indicates recent simultaneous dispersal into east and west Africa. PLoS Negl Trop Dis 7(3):31910. doi:10.1371/journal.pntd.0001910

9. Rico-Hesse R, Harrison LM, Salas RA, Tovar D, Nisalak A, Ramos C, Boshell J, de Mesa MT, Nogueira RM, da Rosa AT (1997) Origins of dengue type 2 viruses associated with increased pathogenicity in the Americas. Virology 230:244–251. doi:10.1006/viro.1997.8504
10. Zehender G, Ebranati E, Shkjezi R, Papa A, Luzzago C, Gabanelli E, Lo Presti A, Lai A, Rezza G, Galli M, Bino S, Ciccozzi M (2013) Bayesian phylogeography of Crimean-Congo hemorrhagic fever virus in Europe. PLoS One 8(11):e79663. doi:10.1371/journal.pone.0079663
11. Olayemi A, Obadare A, Oyeyiola A, Igbokwe J, Fasogbon A, Igbahenah F, Ortsega D, Asogun D, Umeh P, Vakkai I, Abejegah C, Pahlman M, Becker-Ziaja B, Günther S, Fichet-Calvet E (2016) Arenavirus diversity and phylogeography of *Mastomys natalensis* rodents, Nigeria. Emerg Infect Dis 22:687–690. doi:10.3201/eid2204.150155
12. Tsetsarkin KA, Chen R, Yun R, Rossi SL, Plante KS, Guerbois M, Forrester N, Perng GC, Sreekumar E, Leal G, Huang J, Mukhopadhyay S, Weaver SC (2014) Multi-peaked adaptive landscape for chikungunya virus evolution predicts continued fitness optimization in *Aedes albopictus* mosquitoes. Nat Commun 5:1–14. doi:10.1038/ncomms5084
13. Palacios G, Druce J, Du L, Tran T, Briese T, Conlan S, Quan PL, Hui J, Marshall J, Simons JF, Egholm M, Paddock CD, Shieh WJ, Goldsmith CS, Zaki SR, Catton M, Lipkin WI (2008) A new arenavirus in a cluster of fatal transplant-associated diseases. N Engl J Med 358:991–998. doi:10.1056/NEJMoa073785
14. Nei M, Kumar S (2000) Molecular evolution and phylogenetics, 2nd edn. Oxford University Press, Oxford
15. Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O (2011) Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst Biol 60:685–699. doi:10.1093/sysbio/syr041