Chapter 4

Estimation of Evolutionary Dynamics and Selection Pressure in Coronaviruses

Muhammad Munir and Martí Cortey

Abstract

Evolution of coronaviruses is facilitated by the strong selection, large population size, and great genetic diversity within the susceptible hosts. This predisposition is primarily due to high error rate, and limited proofreading capability of the viral polymerase and by recombination. These characteristics make coronaviruses an interesting model system to study the mechanisms involved in viral evolution and the ways viruses adapt to switch host or to gain novel functions. Here we describe the protocol to estimate selection pressures for the spike gene and evolutionary dynamics of bovine coronaviruses.

Key words Coronaviruses, Evolution, Genetics, Emergence, Selection, S gene

1 Introduction

Coronaviruses encode the largest positive sense single-stranded RNA genomes known, ranging from 27 to 31 Kb in length. Although coronaviruses have been shown to possess proofreading ability [1], relatively high mutation rates mean that coronaviruses are one of the most diverse, genetically distinct, and recently emerging groups of viruses. The emergence of these viruses are mainly triggered by the virus evolution which could occur due to high mutational rates, selection pressure on genetic diversity, inter- and intra-host selection, frequency of recombination, and genetic drifts during transmission bottlenecks. Within subfamily *Coronaviridae*, *Alphacoronaviruses*, and *Betacoronaviruses* infect and cause diseases in mammals, whereas *Gammacoronaviruses* are mainly avian specific [2].

Bovine coronaviruses (BCoVs), together with human coronavirus OC43 (HCoV-OC43), equine coronavirus (ECoV), and porcine hemagglutinating encephalomyelitis virus (PHEV), belong to the virus species Betacoronavirus1 of the lineage A of the genus Betacoronavirus [3]. BCoV causes infections both in respiratory and enteric systems in cattle of all ages. Like other coronaviruses,

Helena Jane Maier et al. (eds.), *Coronaviruses: Methods and Protocols*, Methods in Molecular Biology, vol. 1282,
DOI 10.1007/978-1-4939-2438-7_4, © Springer Science+Business Media New York 2015

BCoV exhibit high genetic mutations (one mutation per genome per replication round) [4, 5]. The nucleotide (nt) substitutions per site per year were found to be 1.3×10^{-4}, 6.1×10^{-4}, and 3.6×10^{-4} for RNA-dependent RNA polymerase (RdRp), S, and N genes, respectively [6–8]. Due to their evolutionary potential, BCoVs have been isolated from humans (BCoV-like human enteric coronavirus HECV-4408/US/94) and a recently isolated canine respiratory coronavirus (CRCoV) has also shown a high genetic similarity to Betacoronavirus1 [9, 10].

Taken together, experimental data and mathematical models have reinforced the need for studying coronavirus dynamics and evolution, which could provide bases for effective control measures. Recent availability of quantitative deep-sequencing methodologies has provided data that can be modelled for future prediction of transmission dynamics and to estimate relevant parameters. In this protocol, we used publically available S gene data on BCoV, as prototype coronavirus, and analyzed to predict epidemiological linkage, mutation-prone sites and evolution in the S gene of BCoV. The same protocol is applicable to other genes of the coronaviruses and viruses of other families.

2 Materials

To perform in silico analysis of the S genes of BCoV, the following equipment will be required (*see* **Note 1**):

1. Mac OS X with minimum 2.4 GHz processor and 2 GB RAM.
2. TextEdit (TextWrangler) stable release 1.8 or latest.
3. BioEdit version v7.2.5.
4. MrBayes version 3.2.2 or latest.
5. A Perl script for generating suitable file formats.
6. BEAST version 1.8.0 or latest.
7. BEAUti version 1.7 or latest.
8. Tracer version 1.6 or latest.
9. FigTree v1.2.3 or latest.
10. An appropriate Internet access.

3 Methods

The following procedures are adapted for Mac OS but are equally applicable for other systems.

3.1 Phylodynamics

1. Define objectives (*see* **Note 2**).
2. Construct dataset and label it as BCoV_S genes.fas (*see* **Note 3**).

3. Open the downloaded file (BCoV_S genes.fas) in TextEdit and edit the sequence titles. The sequence titles can be arranged depending upon objective in mind and the availability of downstream analysis tools. One accepted way of labelling the sequence title will be to arrange them in host/isolate_ID/genotype/country/year (accession number). Remove all illegal characters along with empty spaces and replace them with underscore/understrike (_). However, do not remove any greater than signs (>), which will destroy the .fasta format and may require rebuilding of the data set [11]. To do so, use the "Find" and "Replace with" options in the TextEdit, which can be opened with "cmd+F" command in Mac OS X. Save the file before closing the dataset.
4. Open the file in BioEdit and click on Accessory Application -> ClustalW Multiple Alignment. Save the newly opened aligned file and label it as BCoV_S genes_align.fas (*see* **Note 4**).
5. Convert the .fas file (BCoV_S_genes_align.fas) to a .nex file (BCoV_S genes_align.nex): Use either a Perl script (available to freely download at https://github.com/drmuhammadmunir/perl/blob/master/ConvertFastatoPhylip) or using trial version of CodonCode Aligner (www.codoncode.com/aligner/) (*see* **Note 5**).
6. Move BCoV_S_genes_align.nex file into the folder of MrBayes. Detailed description of the program can be found on the webpage of the program (http://mrbayes.sourceforge.net/). Briefly, open terminal and type "mb" to start the MrBayes software (double click on MrBayes application icon in Window). The following instructions should appear:

```
MrBayes v3.2.1 x64

(Bayesian Analysis of Phylogeny)

Distributed under the GNU General Public License

Type "help" or "help <command>" for information on the
     commands that are available.

Type "about" for authorship and general information about
     the program.

MrBayes >
```

7. To execute the file into the program, type "execute <Space>filename" (e.g., execute BCoV_S_genes_align.nex) then press "Enter". The message "Reached end of file" indicates successful execution of the file and the program is ready to run. In any error, either follow the instructions mentioned in the error or rebuilt datasets. The most common error is the presence of illegal characters such as pipeline sign (|), colon (:), semicolon (;) slash/stroke/solidus (/), apostrophe (' '),

quotation marks (‘ ’, “ ”, ' ', ” “), and brackets ([ ], (), { }, < >), among others. Therefore remove these from the fasta file as described before.

8. To set the evolutionary model to the GTR substitution, type “lset nst=6 rates=invgamma” after the MrBayes > prompt then press “Enter”. The message “Successfully set likelihood model parameters” indicates the success in model setup.
9. To set the sample collection (200) from posterior probability distribution, diagnostic calculation every 1,000 generations, and print and sample frequency to 100, type “mcmc ngen=20000 samplefreq=100 printfreq=100 diagnfreq=1000” after the MrBayes > prompt then press “Enter”. Program will start calculating the split frequency depending on the speed of the operating system and the size of the dataset. Note the message “Average standard deviation of split frequencies”. If it is below 0.01 after 2,000 generations, type “yes” after “Continue the analysis? (yes/no)” prompt to set more generations. Continue this until the split frequency drops below 0.01. Once reached, type “no” which leads the users to MrBayes > prompt.
10. To summarize the parameter, type “sump” then press “Enter”.
11. To summarize the tree type “sumt” then press “Enter”. This command will save the tree with extension “nex.con.tre” (i.e., BCoV_S genes_align.nex.con.tre) in the MrBayes folder where the original file (BCoV_S genes_align.nex) was kept. The tree can be opened and annotated in the FigTree.
12. Open the desired file (BCoV_S_genes_align.nex.con.tre) after launching FigTree.
13. Label your sequences by searching your sequence-tag, such as isolate name or country, in the search button when “Taxa” is selected. Similarly, select “Nod” or “Clade” to label the respective items (*see* **Note 6**).
14. After annotation, save your tree using File -> Export Graphics -> PDF (or other desired file format from the list) -> OK path. The resulting file can be used for further editing or for presentation [11].

3.2 Selection Pressures

3.2.1 SNAP

1. To analyze the occurrences of synonymous (dS) and non-synonymous (dN) substitutions in the S gene, use the same fasta file (BCoV_S_genes_align.fas) that was generated for phylodynamics (*see* **Note 7**).
2. Open the SNAP tool freely available at http://www.hiv.lanl.gov/content/sequence/SNAP/SNAP.html and paste the sequence or upload the dataset.
3. Both accumulated (cumulated dN-dS) and per codon (dN-dS) selection sites can be calculated by the generated table of SNAP.

4. Since the selections are calculated on every nucleotide, sites under positive or negative selection can be highlighted (*see* **Note 8**).

3.2.2 Datamonkey

1. Alternatively and to verify the robustness of the data generated by the SNAP, the same alignment can be used to calculate selection pressure using GTR (general time reversible) substitution model on a neighbor-joining phylogenetic tree by the Datamonkey Web server (Freely available at http://www.datamonkey.org/dataupload.php).
2. The program uses the computational engine of the HyPhy package [12] to estimate dN–dS with a variety of evolutionary models and can analyze selection even in the presence of recombination (*see* **Note 9**).

3.3 Evolutionary Dynamics

3.3.1 XML File Generation

1. Within the BEAST package, open BEAUti program (Bayesian Evolutionary Analysis Utility) and import Nexus (BCoV_S_genes_align.nex) or Fasta (BCoV_S_genes_align.fas) file of the data set. Remember to execute the data by "File -> Import Data -> Open".
2. Several parameters of the BEAST run (i.e., the date of the sequences, the substitution model, the rate variation among sites, the length of the MCMC chain) can then be adjusted according to specific need [13] (*see* **Note 10**).
3. Once all desired parameters are set, finally, click on the "Generate BEAST File" to generate .XML file which will be used as input for BEAST analysis.
4. Label the file as BCoV_S_genes_align.xml for consistency.

3.3.2 BEAST Analysis

This is a brief explanation in order to run BEAST program and summarize results using TRACER.

1. Move the .xml files (BCoV_S_genes_align.xml) into the BEAST folder.
2. Open the BEAST program (double-click), a white screen on JAVA environment will appear, wait for several seconds until a second screen appears.
3. Choose the file to analyze in this second screen. Before beginning the analyses enable the "Allow overwriting of log files" option. Then press "Run" and the analysis will begin.
4. After few moments, depending upon the processing capacity of the operating system and the size of the data, the chain will begin to run. There will be seven columns that extend vertically. Every column is one of the parameters that are being estimated; however, the first and the last column are crucial to observe. The first column is the generation being sampled in every moment (every chain has ten million steps) and the last

column shows how many millions of states will be run per hour (remember, ten million steps per chain). Depending on the length of the chain, the length of the sequences, and the number of sequences to be analyzed, it may take variable time to complete the run.

5. Once the chain has run, it is required to store the parameters. Close the BEAST window and open the BEAST folder. Every time a chain is run, two files are generated: .xml file and several ends (.log and .tre). Once the first run is complete, change the name of the .log and .tre files. For example, after the completion of a run for BCoV_S_gene_align.xml, BCoV_S_gene_align.log and BCoV_S_gene_align.tre files will be generated. Rename these two files to BCoV_S_gene_align1.log and BCoV_S_gene_align1.tre.
6. Run the BCoV_S_gene_align.xml at least for two more times (**steps 2–5**).
7. Finally, three different .log files and .tre files will be available labelled as BCoV_S_gene_align1.log, BCoV_S_gene_align2.log, and BCoV_S_gene_align3.log. These three files contain the estimations of the substitution rate that have to be summarized in TRACER.
8. To summarize the run, open the TRACER program and select the option "File" and "Input Trace file", and open the first …1.log file from the folder, followed by the addition of the second (…2.log) file. Finally add the third (…3.log) log file.
9. The estimations of the parameters are viewable in the graphic interface. Select the option "Combined" from the Trace Files (Upper left) and the estimations that will appear on the traces table are the main estimations for all the parameters (*see* **Note 11**). Generally, the desired parameters are:

   Tree Model Root: This is the number of years that passed after the most recent common ancestor (TMRCA). Subtract this number from the most modern date to yield the TMRCA for the dataset.

   Clock Rate: This is directly the rate of evolution in substitution/site/year.

4 Notes

1. This protocol is optimized for Mac OS X; however, all the software packages and tools used here are also available for Windows which can be installed using recommended methodologies. All the software used here are Open Access, which do not require any subscription for any operating systems.

These software packages are only for demonstration purposes, and there may be alternative solutions for the same purpose. The overall time of the data analysis depends upon processing power of the operating system and the number and length of sequences in the dataset.

2. The same phylogenetic tree can be used for different interpretations. Failing to create a proper objective can lead to drawing incorrect conclusions from phylogenetic studies. It is therefore essential to define the objective for the downstream analyses before initiating the study.

3. Construction of datasets depends on the objectives. One of the most common interests of bioinformaticians is to determine the epidemiological linking of the query sequence to that of sequences reported from the world and are available in the public domains. For this purpose, the Basic Local Alignment Search Tool (BLAST) is the most widely used tool, primarily owing to its speed of execution. Search the nucleotide sequences with objective-based keyword such as "Bovine Coronaviruses S gene". Manual editing and investigations of the downloaded sequences are always suggested. Notably, BLAST-Explorer is primarily aimed at helping the construction of sequence datasets for further phylogenetic study, and it can also be used as a standard BLAST server with enriched output. Use BLAST or BLAST-Explorer or other suitable database for construction of datasets.

4. There are different algorithms for DNA sequence alignment with variable degrees of utility. In this protocol, ClustalW was used for simplicity. Any other algorithm can be used depending upon the preferences and interest.

5. Nexus format is required input for MrBayes. Different tools, both online and offline, can be used to generate appropriate nexus output. We have only presented two commonly used and easily achievable methods.

6. Detailed demonstration for tree annotation is described in our earlier publication [11].

7. The file used for phylogenetic analysis may contain all available sequences in the public domain, which increases the size of the file significantly. However, depending upon the objective in mind, the datasets can be modified accordingly. For the larger datasets, the compiled data will be emailed to the email address provided once ready. This is also important to keep a record for future use.

8. The cut point is calculated to be zero. All sites showing cumulated dN-dS values above 0 are under positive pressure whereas values below 0 are under negative pressure.

9. The nature of parameter selection and interpretation is complex and is beyond the scope of this protocol. Please consult developers' published report for thorough understating of the concepts and applications [12].
10. The parameters of the BEAST run are crucial and can determine the nature of output and may heavily influence the results. However, normally the default parameters are used [13].
11. When summarizing the BEAST results, do not use the mean as it appears in the output (this is the arithmetic mean); instead use the geometric mean that appears in the summary statistic table (Right Upper). From a methodological point of view this is much more correct (Click on every parameter in the Tracer table and the summary statistic table for every parameter will change).

Acknowledgement

The work at our laboratories is supported by the grant from Biotechnology and Biological Sciences Research Council (BBSRC) through Institute Strategic Programme Grant (BB/J004448/1).

References

1. Minskaia E, Hertzig T, Gorbalenya A et al (2006) Discovery of an RNA virus 3′→5′ exoribonuclease that is critically involved in coronavirus RNA synthesis. Proc Natl Acad Sci U S A 103:5108–5113
2. Jackwood MW, Hall D, Handel A (2012) Molecular evolution and emergence of avian gammacoronaviruses. Infect Genet Evol 12: 1305–1311
3. de Groot RJ, Baker SC, Baric R et al (2012) Coronaviridae. In: King AMQ et al (eds) Virus taxonomy: classification and nomenclature of viruses: ninth report of the international committee on taxonomy of viruses. Elsevier Academic Press, Oxford, pp 806–828
4. Drake JW, Holland JJ (1999) Mutation rates among RNA viruses. Proc Natl Acad Sci U S A 96:13910–13913
5. Moya A, Holmes EC, Gonzalez-Candelas F (2004) The population genetics and evolutionary epidemiology of RNA viruses. Nat Rev Microbiol 2:279–288
6. Vijgen L, Keyaerts E, Lemey P et al (2006) Evolutionary history of the closely related group 2 coronaviruses: porcine hemagglutinating encephalomyelitis virus, bovine coronavirus, and human coronavirus OC43. J Virol 80:7270–7274
7. Woo PC, Lau SK, Lam CS et al (2012) Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. J Virol 86:3995–4008
8. Erles K, Shiu KB, Brownlie J (2007) Isolation and sequence analysis of canine respiratory coronavirus. Virus Res 124:78–87
9. Bidokhti MR, Tråvén M, Krishna NK et al (2013) Evolutionary dynamics of bovine coronaviruses: natural selection pattern of the spike gene implies adaptive evolution of the strains. J Gen Virol 94:2036–2049
10. Zhang XM, Herbst W, Kousoulas KG et al (1994) Biological and genetic characterization of a hemagglutinating coronavirus isolated from a diarrhoeic child. J Med Virol 44: 152–161
11. Munir M (2013) Bioinformatics analysis of large-scale viral sequences: From construction of data sets to annotation of a phylogenetic tree. Virulence 4:97–106
12. Pond SLK, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679
13. Drummond AJ, Suchard MA, Xie D et al (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29: 1969–1973