Visualizing genetic signatures of foot-and-mouth disease virus associated with host factors

Garabed, R.B[1]; Perez, A.M[1, 2]; Knowles, N.J.[3]; Valarcher, J.F.[4]; Thurmond, M.C.[1]
[1] FMD Modeling and Surveillance Laboratory, University of California, Davis; [2] CONICET/INTA Balcarce, Argentina; [3] Pirbright Institute for Animal Health; [4] IVI-Animal Health. Lärkbacken, Uppsala, Sweden

Abstract

Identifying specific directed genetic variation in a highly variable virus like foot-and-mouth disease virus (FMDV) is of great importance for understanding and combating the disease. Using Bayesian statistics with semi-informative prior information, and using GIS software to visualize the resulting genetic signatures, significant differences in nucleotide sequences among several FMDV subpopulations can be identified. Probabilities of a nucleotide being present at each locus are calculated by a Bayesian multinomial-response mixed-effects logistic regression to account for the non-random collection and sequencing procedures. The estimated nucleotide probabilities at each locus were used to construct filters to remove loci that provide no information about differences in the genome among isolates. The filtered probabilities of nucleotides for the population of viruses are plotted as a surface in ArcScene© and the probabilities for subpopulations (with 95% probability intervals) are plotted as bars. The plot allows for visualization and identification of loci that have significantly different nucleotide probabilities based on host factors. An application of the methods is illustrated using a set of 339 aligned, FMDV, serotype-O, VP1 sequences. These methods may be useful in identifying locations in the genome for study by molecular biologists and for identifying emergence of atypical FMDV strains.

Introduction

Foot-and-mouth disease virus (FMDV) is notable in its ability to infect a variety of host species, persist in many geographic regions, and vary in its genome structure. Some FMDV genetic variability may be adaptation to different host species and environmental conditions, but identifying the specific nucleotide changes in a highly variable genome is difficult. Understanding FMDV genetic variability is complicated further by inconsistencies in choice of virus isolates collected for sequencing and in associated host demographic data. A major limitation to applying statistical methods to genetic epidemiology is the inability to associate hypothesized/putative exposure or host factors with mutation. Reasons for this include the limited number of available nucleotide sequences with associated epidemiologic information, difficulty in determining the appropriate response to model, and a non-random process by which sequenced viruses have been sampled. In the present study, we provide a method of quantifying and visualizing host-associated risk factors for mutation that is applicable to viruses like FMDV for which a variety of sequences exist with associated epidemiologic information.

Objective

To quantify and to display relationships between specific mutations in FMDV and hypothesized host factors.

Methods

General Approach

A Bayesian multinomial-response, mixed-effects logistic regression model was fit to nucleotide sequence data of FMDV serotype-O isolates that spanned a wide variety of isolation times, locations, and host species.

Data

Aligned sequences of the capsid encoding region, covering 636 nucleotides and referred to as the VP1 gene, of 339 FMDV serotype-O isolates collected from around the world over the years 1948-2003 were provided by N.J. Knowles and J.F. Valarcher. Information was available for country of origin, isolation year, and species from which the virus was isolated. Species from which the viruses were isolated were bovine (n=199), porcine (n=67), ovine or caprine (n=16), buffalo (n=12), and other (n=45). The "other" designation included exotic ruminants and animals of unknown or unrecorded species. These sequences were assigned to clusters or networks of viruses, based on an expert assessment of the topotype, submitter, location of isolation, and date of isolation, so that isolates that were considered to be related were treated as such in the model.

Model

The data were used to construct a Bayesian multinomial-response mixed-effects logistic regression, which was fit using WinBUGS© (Imperial College & MRC, UK). The covariates that were believed to be associated with mutations were host species, continent of origin, and year of isolation. Additionally, a random cluster effect was used to account for possible repeated sampling of some viral lineages that were based on expert opinion. The response modeled was the probability of finding a specific nucleotide base at a site (locus) in the VP1 gene. Each site was treated as entirely independent of the other sites. The probabilities for each site were calculated using the same semi-informative priors based on expert opinion. The priors were placed on the probability of each base occurring at any random site in the FMDV VP1 gene and, thus, were designed to partially account for the multiple comparisons used in the analysis.

Display

The probabilities for finding each of the four bases, an unknown base, or a gap at each site were calculated from the model (with 95% probability intervals) for different species, continents, and years. Sites with no variation were defined as those with 98% or more probability of a specific base. Those sites that had essentially no variation or only variation among two bases that would produce the same protein sequence (synonymous mutations) were dropped from the display to reduce clutter. Sites with only synonymous mutations were defined as sites with less than 2% probability given to non-synonymous mutations. The remaining sites were plotted in ArcScene™ (ESRI®, Redlands, CA) with a baseline (bovine for example) of probabilities as a surface and a specific population (buffalo for example) plotted as a bar. ArcScene™ allows the user to zoom and to change the angle of view of these surfaces so that the whole gene can be explored with ease.

Results

Figure 1 shows a preliminary example of a plot (where synonymous mutations have not yet been dropped and 95% intervals have not been included) produced by this method with the probabilities for buffalo shown in magenta versus the baseline (bovine) shown in grey for the 318th to 345th positions in the VP1 gene. Here significant mutations at sites 318 (G), 321 (C), 342 (C), and 345 (A) stand out high above the surface. Thus, buffalo appear to be more likely than cattle to have a G at site 318, a C at sites 321 and 342, and an A as opposed to a G at site 345.

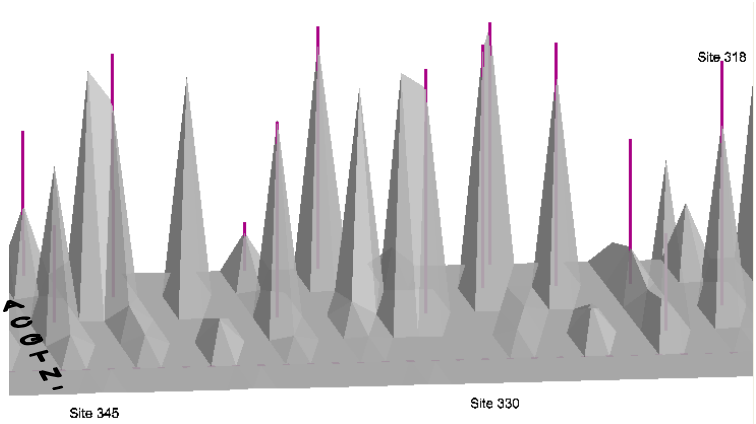

Figure 1 Predicted probabilities of each nucleotide base (A, C, G, T, unknown, and gap from farthest to closest) for bovine (grey) versus buffalo (magenta) at the 345th to 318th positions out of 636 in the VP1 gene

Discussion

Results of the method presented here provide a simple way to view non-synonymous mutations in the FMDV genome that appear to be significantly related to a risk factor. The risk factors that could be tested in this example were limited by the quality of the epidemiologic data that were collected along with the isolates submitted for sequencing. Nevertheless, this example provides some interesting results. It is important that the results of this model be tested for their biological meaning, i.e. do the amino acid changes resulting from these mutations decrease host recognition and elimination of the virus. Additionally, it is important to note that the assignment of clusters, presence of missing data, and choice of priors affect the results of this model. For example, if 50% of the "other" host species category is assigned to the porcine versus buffalo species, the mutations found to be significant in these two categories would both change. However, model certainty would be expected to improve as reporting of epidemiologic data improves and as more and representative samples are sequenced.