Available online at www.sciencedirect.com
ScienceDirect
Infection, Genetics and Evolution 8 (2008) 130–138
Infection, Genetics and Evolution
www.elsevier.com/locate/meegid

Nucleotide sequence polymorphism in circoviruses

Austin L. Hughes [a,*], Helen Piontkivska [b]

[a] Department of Biological Sciences, University of South Carolina, Coker Life Sciences Building, 700 Sumter St. Columbia, SC 29208, United States
[b] Department of Biological Sciences, Kent State University, Kent, OH 44242, United States

Received 13 August 2007; received in revised form 2 November 2007; accepted 7 November 2007
Available online 17 November 2007

Abstract

Analysis of nucleotide diversity within six species of circovirus showed consistently stronger purifying selection at nonsynonymous sites in the *rep* gene than on those in the *cap* gene. In addition, synonymous nucleotide diversity in the *rep* gene was significantly lower than that in the *cap* gene, suggesting functional constraint even at synonymous sites in *rep*, which was associated in all six species with strongly negative AT-skew. Of the six virus species examined, four species showed evidence of ongoing purifying selection at nonsynonymous polymorphic sites in the *rep* gene, indicating the presence of slightly deterious nonsynonymous variants in these populations. The *rep* gene of porcine circovirus 2 (PCV2) was unique, however, in showing a strong excess of rare nonsynonymous polymorphisms. The excess of rare nonsynonymous polymorphisms suggests a prolonged population bottleneck in PCV2, allowing slightly deleterious mutations to accumulate, followed by a population expansion during which selection to remove these variants has increased in effectiveness. Such a population history is consistent with the epidemiological evidence of a recent worldwide spread of PCV2.

Keywords: Circovirus; Population bottleneck; Purifying selection; Virus evolution

1. Introduction

The circoviruses are a family of viruses with a circular, single-stranded DNA genome, members of which have been found to infect a number of species of birds and mammals. The circovirus genome is remarkably compact, typically including just two major genes: *rep* (encoding the replicase protein) and *cap* (encoding the capsid protein). However, a number of alternative products of the *rep* reading frame have been reported (Mankertz and Hillenbrand, 2001; Cheung, 2003a,b,c). There are two closely related circoviruses infecting pigs, porcine circovirus 1 (PCV1) and porcine circovirus 2 (PCV2). PCV1 was first discovered as a contaminant of cultured pig kidney cell line (Allan and Ellis, 2000). PCV1 is non-pathogenic, and is believed to be widespread among pigs worldwide. PCV2, on the other hand, is an emerging virus of pigs that is associated with certain disease syndromes, especially postweaning multisystemic wasting syndrome (PMWS), characterized by wasting, dyspnea, and enlarged lymph nodes in pigs around 5–12 weeks of age (Allan and Ellis, 2000). While PCV2 appears to be necessary for the appearance of PMWS, it is apparently not sufficient; and infection with other viruses may be required for full development of the syndrome. PMWS was first identified in Western Canada in 1991 and was subsequently reported elsewhere in North America and in Europe, leading to the hypothesis that the spread of PCV2 has been recent (Meehan et al., 1998).

Because of the close relationship between PCV1 and PCV2, comparison of these viruses can potentially provide insights into the evolutionary origin of PCV2 and the molecular basis of pathogenesis, including the role of natural selection (Olvera et al., 2007). Natural selection is an important factor in understanding genomic evolution, including both positive selection (favoring adaptive mutations) and purifying selection (acting to eliminate deleterious mutations; Hughes, 1999). Viruses have provided some of the best-documented examples of positive Darwinian selection at the DNA sequence level, particularly selection exerted by the host immune system favoring the evasion of immune recognition (Allen et al., 2000; Evans et al., 1999; Fitch et al., 1991; Hughes et al., 2005; O'Connor et al., 2004; Moore et al., 2002; Seibert et al., 1995). However, as is the case with cellular organisms, the predominant form of natural selection on viral genomes is

* Corresponding author. Tel.: +1 803 777 9186; fax: +1 803 777 4002.
E-mail address: austin@biol.sc.edu (A.L. Hughes).

1567-1348/$ – see front matter
doi:10.1016/j.meegid.2007.11.001

purifying selection (Hughes, 2007a; Hughes and Hughes, 2005; Hughes et al., 2007; Jerzak et al., 2005; Nei, 1983; Pybus et al., 2007; Saitou and Nei, 1986; Suzuki and Gojobori, 1997). Because purifying selection acts most strongly on genomic regions that are functionally important, the patterns of purifying selection can be used to identify regions that are least likely to change over the course of evolution. In the case of viruses, knowledge of evolutionarily conserved regions can aid in the design of vaccines and other therapeutic agents because it can help predict the likelihood of evolution of resistant viral genotyes (Brown et al., 2007; Haydon et al., 2001; Slobod et al., 2005; Storgaard et al., 1999).

The strongest evidence for the predominance of purifying selection is the observation that the number of synonymous nucleotide substitutions per synonymous site (d_S) substantially exceeds the number of nonsynonymous substitutions per nonsynonymous site (d_N) in most genes (Kimura, 1977), a pattern observed in viruses as in cellular organisms (Saitou and Nei, 1986). This pattern occurs because most nonsynonymous mutations are disruptive to protein structure and thus tend to be eliminated by purifying selection (Kimura, 1977). A strongly deleterious mutation may be eliminated immediately; for example, if the mutation has a lethal effect. However, the effectiveness of purifying selection on slightly deleterious mutations depends on the effective population size, being more effective in larger populations (Ohta, 1973, 1976, 2002).

During a population bottleneck, slightly deleterious mutations can drift to relatively high frequencies and may even become fixed. A characteristic signature of a species that has undergone expansion following a prolonged bottleneck is an excess of rare nonsynonymous polymorphic variants that show evidence of ongoing purifying selection (Hughes et al., 2003; Hughes and Hughes, 2007). These rare nonsynonymous variants represent slightly deleterious mutations that drifted to relatively high frequencies during the bottleneck, when selection against them was ineffective. As the population expands, selection against such slightly deleterious variants becomes increasingly effective, leading to a reduction in allelic frequencies.

Here we analyze nucleotide sequence polymorphism within six circovirus species in order to examine in a comparative framework the patterns of natural selection on the two major protein-coding genes and on intergenic regions. Using an extensive sample of complete genome sequences of PCV2, we analyze the patterns of synonymous and nonsynonymous polymorphism in order to test the hypothesis that this virus has undergone a rapid population expansion following a prolonged bottleneck. We also tested for the role of additional factors, such as nucleotide content and the coding of alternative products (in the case of *rep*), in accounting for patterns of sequence conservation.

2. Methods

2.1. Alignment and phylogenetic analysis

The complete genomes of the following six virus species were obtained from the NCBI data base: beak and feathers disease virus (BFDV), 21 genomes; columbid circovirus (CoCV), 5 genomes; goose circovirus (GCV), 23 genomes; muscovy duck circovirus (MuDCV), 6 genomes; porcine circovirus 1 (PCV1), 22 genomes; and porcine circovirus 2 (PCV1), 215 genomes. For accession numbers of the sequences, see Supplementary Fig. S1. Only genomes with full *rep* and *cap* open reading frames were used. Within each of the six viral species, non-coding and coding regions were aligned using the CLUSTAL X program (Thompson et al., 1997). Coding sequences were aligned at the amino acid level and the alignment imposed on the DNA sequence. In all pairwise comparisons among a set of sequences, any site at which the alignment postulated a gap in any of the sequences compared was excluded from all pairwise comparisons.

The Rep protein of circoviruses shows homology to that of the nanoviruses (Niagaro et al., 1998; Hughes, 2004); therefore, the amino acid sequences of the Rep protein were aligned with a Rep protein sequence from banana bunchy top virus (BBTV), which was used as an outgroup to root a phylogenetic tree of Rep sequences (Supplementary Fig. S1). Phylogenetic trees were also constructed from nucleotide sequences of both *rep* and *cap* genes of the closely related species PCV1 and PCV2. The following methods were used for phylogenetic reconstruction: (1) the neighbor-joining (NJ) method (Saitou and Nei, 1987); (2) and the Bayesian method (Huelsenbeck and Ronquist, 2001). The NJ trees of amino acid sequences was constructed on the basis of the equal-input model (Kumar et al., 2004). The Bayesian tree of amino acid sequences was reconstructed using the JTT + Γ model, which allows for rate variation among sites (Rodriguez et al., 1990). All parameters were estimated from the data. Four chains were run for 2,500,000 generations, and trees were sampled every 100 generations. Bayesian posterior probabilities were inferred from the last 5000 sampled trees. The reliability of branching patterns in MP and NJ trees was estimated by bootstrapping (Felsenstein, 1985); 1000 bootstrap pseudo-samples were used. The *rep* and *cap* sequences of the common ancestor of the 215 PCV2 genomes were reconstructed by the maximum parsimony method on the basis of NJ trees using PCV1 sequences as an outgroup.

2.2. Nucleotide sequence diversity

The number of synonymous substitutions per synonymous site and the number of nonsynonymous substitutions per nonsynonymous site were estimated by Nei and Gojobori (1986) method, using the MEGA3 software (Kumar et al., 2004). Within each of the six virus species, the mean for all pairwise comparisons of the number of synonymous substitutions per synonymous site provided an estimate of nucleotide diversity at synonymous sites (π_S); and the mean for all pairwise comparisons of the number of nonsynonymous substitutions per nonsynonymous site provided an estimate of nucleotide diversity at synonymous sites (π_N) (Nei and Kumar, 2000). The number of nucleotide substitutions per site (d) in non-coding regions was estimated by Jukes and Cantor's model; the nucleotide diversity (π) in non-coding regions