Virology 422 (2012) 195–204

Contents lists available at SciVerse ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Identification and classification of endogenous retroviruses in the canine genome using degenerative PCR and *in-silico* data analysis

Haiin Jo [a], Hojun Choi [a], Min-Kyeung Choi [a], Ning Song [a], Jin-Hoi Kim [a], Jae-Wook Oh [b], Kunho Seo [c], Han Geuk Seo [a], Taehoon Chun [d], Tae-Hun Kim [e], Chankyu Park [a,*]

[a] Department of Animal Biotechnology, Konkuk University, Hwayang-dong, Kwangjin-gu, Seoul, South Korea
[b] Division of Animal Life Science, Konkuk University, Seoul, South Korea
[c] College of Veterinary Medicine, Konkuk University, Seoul, South Korea
[d] College of Life Sciences and Biotechnology, Korea University, Seoul, South Korea
[e] Animal Genomics and Bioinformatics Division, National Institute of Rural Development Administration Suwon, South Korea

ARTICLE INFO

Article history:
Received 1 July 2011
Returned to author for revision
15 August 2011
Accepted 11 October 2011
Available online 14 November 2011

Keywords:
Endogenous retrovirus
Canine
Genome
Pro
Pol
ERV

ABSTRACT

Pooled genomic DNA from 10 dogs was subjected to polymerase chain reaction with primers targeting the retroviral *pro/pol* region. Sequence analysis of 120 clones obtained by PCR revealed 81 of retroviral origin. Subsequent analysis of the dog genome (CanFam 2.0) by BLAST investigation using degenerate PCR products and previously identified retroviral sequences permitted the identification of additional retroviral γ and β sequences. A phylogenetic analysis using the retroviral protease (PR) and reverse transcriptase (RT) sequences in the dog genome resulted in identification of 17 γ and 7 β families. In addition, we also identified 167 spuma-like ERV elements from CanFam 2.0 based on sequence homology to murine (Mu)ERV-L and human (H)ERV-L. Our results could contribute to the understanding of the influence of retroviruses in shaping the genome structure and altering gene expression by providing quantitative and locational information of ERV loci and their diversity in the dog genome.

Introduction

Retroviruses can be broadly classified into two groups: simple (alpha, beta, gamma and epsilon retroviruses) and complex (lenti-, delta- and spuma-viruses). Only simple retroviruses are capable of being endogenous in their host genome with the exception of spuma-viruses (Weiss, 2006). Endogenous retroviruses (ERVs) integrate into the host germ line in a Mendelian order (Gifford and Tristem, 2003). Their typical structure consists of *gag* (viral core proteins), *pro–pol* (viral enzymes), and *env* (envelope proteins) genes with long terminal repeats (LTRs) at their 5′ and 3′ termini (Patience et al., 2001).

It has been proposed that ERVs co-evolved with their host genome (Britten, 1996; McDonald, 1993) and similar retroviral sequences in diverse species suggest cross-species transmission events (Benit et al., 2001). As a result, ERVs carry a significant disposition of numerous point mutations, deletions, and insertions over time, and thus act as a 'fossil record' for understanding retroviruses and the co-evolution of their host genome (Coffin et al., 1997). Therefore, ERV-originated sequences are sometimes difficult to recognize due to severe deterioration as a result of accumulated mutations.

In most cases, ERVs no longer maintain their activity, but ERV insertions produce diverse effects on their host genome including altered gene expression by causing ectopic recombination (Hughes and Coffin, 2005), premature polyadenylation (Palmarini et al., 2002), or aberrant splicing (Maksakova et al., 2006). Therefore, the changes in the genome caused by ERVs can result in diseases in organisms (Maksakova et al., 2006; Ryan, 2004).

ERVs have been studied in diverse species including mammals (Tristem, 1996; Tristem et al., 1996), chickens (Dunwiddie et al., 1986), reptiles, amphibians, and fish (Herniou et al., 1998). In mammals, detailed genome level characterizations of ERVs were carried out in several species including humans (Lower et al., 1996), chimpanzees (Polavarapu et al., 2006), swine (Klymiuk et al., 2002; Patience et al., 2001), sheep (DeMartini et al., 2003; Klymiuk et al., 2003), and cattle (Garcia-Etxebarria and Jugo, 2010; Xiao et al., 2008b). From these studies, three classes of ERVS were reported: Class I (γ [C-type]), Class II (β [B/D-type]), and Class III (spuma-like).

* Corresponding author at: Department of Animal Biotechnology, Konkuk University, Hwayang-dong, Kwangjin-gu, Seoul 143–701, South Korea. Fax: +82 2 457 8488.
E-mail addresses: hi458@konkuk.ac.kr (H. Jo), shukle2@konkuk.ac.kr (H. Choi), sirius1029@nate.com (M.-K. Choi), funny-sonia@163.com (N. Song), jhkim541@konkuk.ac.kr (J.-H. Kim), ohjw@konkuk.ac.kr (J.-W. Oh), bracstu3@konkuk.ac.kr (K. Seo), tchun@korea.ac.kr (T. Chun), thkim63@korea.kr (T.-H. Kim), chankyu@konkuk.ac.kr (C. Park).

0042-6822/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.virol.2011.10.010

The dog is an important animal species as a major companion animal (Clutton-Brock, 1995; Schwartz, 1997) and disease model (Patterson et al., 1982; Patterson et al., 1988; Wayne and Ostrander, 1999). The current dog genome assembly, CanFam 2.0, covers 98% of the entire canine genome with 2400 Mb of assembled sequences (Lindblad-Toh et al., 2005), and the brief estimation of ERV sequences in the dog genome have been attempted (Blikstad et al., 2008; Lindblad-Toh et al., 2005). Very recently, Martínez Barrio et al. (2011) reported a comprehensive analysis on the complexity and integration pattern of canine endogenous retroviruses (CfERVs) using an ERV analysis program. The discovery and detailed characterization of ERVs in the canine genome are important to improve annotation of repetitive elements and facilitate an understanding of the role of ERVs in modification of the genome (Jern et al., 2005; Lower et al., 1996).

A limited number of ERV sequences were initially identified from experimental approaches including the analysis of cDNA-library clones (La Mantia et al., 1991). More recently, a degenerate polymerase chain reaction (PCR) method using specific primers for the conserved active site motifs of protease (PR) and reverse transcriptase (RT) has been successfully employed to study diverse ERVs in the genome (Akiyoshi et al., 1998; Klymiuk et al., 2002; Xiao et al., 2008b).

In this study, we first amplified the conserved *pro*/*pol* region of canine ERVs using degenerate PCRs and analyzed the products after cloning and sequencing. Subsequent *in-silico* searches of ERV sequences against the reference dog genome using the information derived from the degenerate PCRs and previously reported retroviral sequences resulted in a genome level evaluation of ERV elements which we classified into individual ERV families on the basis of *pro*/*pol* sequence diversity.

Results

Identification of ERV gamma (γ) elements using degenerate PCR

Degenerative PCR using *pro*/*pol* specific primers and dog genomic DNA and subsequent cloning generated a total of 120 candidates with possible PR and/or RT containing sequences. Sequence analysis of the clones revealed 86 (71.7%) with a length of 0.6 to 1.0 kb having significant sequence similarity (40–97%) to previously identified γ retroviruses (data not shown). Among clones of retroviral origin, 10 clones of five different pairs had repetitious sequences, showing a result similar to previous studies (Patience et al., 2001; Xiao et al., 2008b). Therefore, we identified 81 unique sequences with retroviral origins using degenerate PCR.

The clones (or elements in *in-silico* analysis) having more than 80% nucleotide sequence identity were clustered as a family into four groups with 48, 25, 6, and 2 clones each. The most abundant group was designated as *Canis familiaris* endogenous retrovirus (CfERV) γ1, followed by γ2, γ3, and γ4. Their phylogenetic relationship showed the same topology as the order of clone abundance, thus the genetic distance between CfERV γ1 and γ4 was furthest (Supplementary Fig. 1A). The sequence having the fewest nonsense mutations and being most closely related to potentially infectious ERVs was determined as the representative for each family. Phylogenetic analysis was performed together with previously identified γ- and β- retroviruses. All sequences from degenerate PCR were clustered with gammaretroviruses and no clones clustered with betaretroviruses (Supplementary Fig. 1B).

The CfERV γ2 sequences were relatively shorter (557–594 base pair [bp]) than others (γ1, γ3, and γ4) which ranged from 605 to 946 bp. Subsequent analysis showed that CfERV γ2 contained the PR region but lacked most RT sequences. However, CfERV γ2 clones should contain at least a part of the RT specific motif sequence to be amplified by *pro*/*pol* specific primers, which was supported by confirming the presence of both PR and RT sequences in the primer site and RNaseH sequences from the sequencing results of γ2 clones (data not shown).

In-silico analysis of dog ERVs from CanFam 2.0 detecting 176 ERV γ and 8 β elements

Basic Local Alignment Search Tool (BLAST) analysis using 81 *pro*/*pol* sequences from degenerative PCR through a NCBI blastn suite, BLAST Dog Sequences, identified 160 loci with ERV γ elements (Supplementary Table 1). To identify β family elements, we performed BLAST analysis in the same way using nine previously identified betaretroviruses (TvERV, SRV, JSRV, OERV β1, OERV β2, OERV β3, HERV-K, MPMV, PERV β5) of several species. Although we used diverse β retroviral sequences, we only found 8 BLAST matches from the entire canine genome (Supplementary Table 2), indicating a limited presence of ERV β elements in the dog genome. Since ERVs are a member of long terminal repeat (LTR) retrotransposons, we tried to identify full-length LTR retrotransposons with both 5′ and 3′ LTRs. Although 557 putative full-length LTR retrotransposons were identified from the LTR_STRUC (McCarthy and McDonald, 2003) analysis, only 39 were confirmed as *pro*/*pol* containing elements (data not shown). Most sequences were retrotransposons with LTRs at both the 5′ and 3′ ends without any significant matches to retroviral sequences. We also attempted to identify ERV sequences using three reported ERV sequences of 7000 bp of *Canis familiaris* origin from the Repbase (http://www.girinst.org.repbase/). We combined all ERV sequences from different approaches, removed redundant or severely mutated sequences including large deletions or insertions which might cause problems in constructing a phylogenetic tree, and chose sequences which contained the *pro*/*pol* region and mapped to different locations in CanFam 2.0. Finally, 173 ERV γ and 8 β elements were identified from the current dog genome assembly (Supplementary Tables 3 and 4). Each element was analyzed for chromosomal coordinates within the genome assembly, element sizes (bp) and length, and sequence identity between 5′ and 3′ LTRs. The size of ERV elements was estimated by counting the number of nucleotides between the start of the 5′ LTR and the last nucleotide of the 3′ LTR. Thus, element length was defined only for those (71.8%) with recognizable LTR sequences at both ends.

In addition, we evaluated 407 CfERV elements reported by Martínez Barrio et al. (2011) according to our ERV determination criteria which were described in Materials and methods. A significant number of CfERV elements defined by them did not meet our criteria and were excluded from our result. Consequently, three γ elements were additionally added to PR/RT containing CfERVs, resulting in the identification of 176 γ and 8 β ERV elements from the dog genome (Supplementary Tables 3 and 4).

A total of 74.4% (n = 131) of ERV γ and 25% (n = 2) of β elements had two recognizable LTRs, while 4.5% (n = 8) of γ elements were solo LTR ERVs; therefore, we were not able to identify LTR sequences from 21.0% (n = 37) and 75% (n = 6) of γ and β elements, respectively. The average size of ERV elements was 8427 bp (range 8259 to 8594) for CfERV β and 6380 bp (3555 to 18462) for γ elements. These results are consistent with those of previous studies from other species (Bowen and McDonald, 1999; Jern et al., 2005; McCarthy and McDonald, 2004; McCarthy et al., 2002). The large variation in size among γ elements was due to the presence of nucleotide deletions and insertions within the elements. For example, the element *Der10-5* (CfERV γ10, 3979 bp) had Gag and PR but lacked RT between 5′ and 3′ LTRs (Supplementary Table 6). In contrast, we found insertions of both transposase and non-annotatable nucleotide sequences within the element *Der18-7* (CfERV γ10, 14434 bp).

To estimate the total amount of ERV sequences in the dog genome, we simply multiplied the copy numbers of β and γ elements by the average size of element length for each genus because the exact boundary of ERV elements with missing or solo LTRs was uncertain.