Gene 548 (2014) 14–21

Contents lists available at ScienceDirect

Gene

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

Genomic environment and digital expression of bovine endogenous retroviruses

Koldo Garcia-Etxebarria [a], Begoña M. Jugo [a,*]

[a] Genetika, Antropologia Fisikoa eta Animalien Fisiologia Saila, Zientzia eta Teknologia Fakultatea, Euskal Herriko Unibertsitatea (UPV/EHU), 644 Postakutxa, E-48080 Bilbao, Spain

ARTICLE INFO

Article history:
Received 25 September 2013
Received in revised form 8 February 2014
Accepted 23 June 2014
Available online 26 June 2014

Keywords:
ERVs
Bos taurus
Insertions
Expression

ABSTRACT

In this work, we attempt to identify the location and expression of bovine ERVs and the nature of nearby genes. As many as 1610 bovine genes contain a bovine ERV (BoERV) inserted into their introns. Most of these BoERVs present an antisense orientation, which could be a consequence of the detrimental effects of the sense insertion. Based on the overrepresentation of Gene Ontology terms, we determined that some genes located in the vicinity of BoERVs are related to viral response and chromatin assembly. In addition, we identified some genes that belonged to IFNs that are inserted in or between BoERVs, pointing out a possible role of BoERVs in some gene duplication. Based on EST database mining, significant expression was detected for BoERV genes in the thyroid and in embryos only. Transcription factor binding sites for Rxra, a steroid and thyroid hormone receptor, were detected in three of the most expressed BoERVs in the thyroid glands.

1. Introduction

Endogenous retroviruses (ERVs) are the remains of the integration of exogenous retroviruses into a host genome (Tristem, 2000). Due to their pathogenic characteristics, ERVs are usually inactivated or eliminated from the host genome (Gifford and Tristem, 2003), but some of them remain integrated in the genome and can also confer benefits on the host.

ERVs may keep their functionality and expression in some animal tissues and at different developmental stages (Jern and Coffin, 2008). For instance, the expression of one type of ERV (enJSRVs) has been detected in ovine embryos as a defense against its exogenous counterpart (Spencer et al., 2003). Expression of endogenous retroviruses has been also observed in some types of cancer and autoimmune diseases (Perron et al., 2009; Prusty et al., 2008). ERV expression has been analyzed experimentally using qPCR (Bittmann et al., 2012; Tarlinton et al., 2013), microarrays (Pérot et al., 2012; Seifarth et al., 2005) and, more recently, transcriptome sequencing analysis (Bolisetty et al., 2012; Brown et al., 2012). Alternatively, ERV expression can be indirectly inferred by searching in silico EST databases (Deloger et al., 2009; Stauffer et al., 2004), which contain a tremendous amount of publicly available data.

Few studies have analyzed the interplay between ERVs and genes in the host genome. The contribution of all transposable elements (TEs) to the host's genetic structure and evolution has been analyzed in a small number of mammalian species, for example in cattle and mice (Almeida et al., 2007; DeBarry et al., 2006), but specific studies of the contribution of LTR-retroelements, where ERVs are classified, are more limited. Some gene/LTR retrotransposon associations have been described in two species of invertebrates, *Caenorhabditis elegans* (Ganko et al., 2003) and *Drosophila* (Ganko et al., 2006), primarily involving genes related to development and behavior (Ganko et al., 2006).

The location itself of transposable elements, including ERVs, could affect gene expression (Hollister and Gaut, 2009). In addition, some studies have shown that the effects of the ERVs on the expression of nearby genes are due to regulatory elements (such as promoters or enhancers) carried in their LTRs (Jern and Coffin, 2008).

Bos taurus is a species of interest because of its economic importance worldwide, and it is the first ruminant whose genome has been sequenced (Elsik et al., 2009). The estimated size of the bovine genome is 2.87 Gbp; it is estimated that 27% of the genome is composed of TEs. Previous to the release of the whole genome sequencing, Almeida et al. (2007) analyzed the location of, different TE-derived sequences in cattle and detected many of them in gene introns and, to a lesser extent, in exons and promoter regions. LTR elements (including ERVs) in particular showed a significant antisense insertion bias in bovine gene introns (Almeida et al., 2007). Based on the cattle Btau_3.1 genome version, bovine ERVs were thoroughly characterized (Garcia-Etxebarria and Jugo, 2010) and 86 complete retroviral regions were described. The location of the characterized elements and their expression levels could provide clues about their role in cattle genetics and physiology.

Abbreviations: ERV, endogenous retrovirus; BoERV, bovine endogenous retrovirus; LTR, long terminal repeat; EST, expressed sequence tag; INF, interferon; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; MHC, major histocompatibility complex.

* Corresponding author.
E-mail addresses: koldo.garcia@ehu.es (K. Garcia-Etxebarria), begonamarina.jugo@ehu.es (B.M. Jugo).

http://dx.doi.org/10.1016/j.gene.2014.06.048
0378-1119/

In this study, the insertion locations of the previously detected bovine ERVs, the insertion direction, and the nature of the neighboring genes have been analyzed; Moreover, an indirect study of the expression of these bovine ERVs by EST database mining has been carried out.

2. Materials and methods

This study used the Btau_3.1 version of the *B. taurus* genome (Elsik et al., 2009), including the annotation of this genome version and the genes described therein, as well as the bovine endogenous retroviruses (BoERVs) detected in this genome version by Garcia-Etxebarria and Jugo (2010) using three different methods (BLAST, LTR_STRUC and Retrotecto©).

2.1. Characterization of genes near BoERVs

To detect genes located near selected bovine endogenous retroviruses (BoERVs) (Garcia-Etxebarria and Jugo, 2010), the described bovine genes and BoERVs were connected by in-house scripts written in PHP. This analysis was performed in two ways: (i) using all BoERVs detected by three different methods (BLAST, LTR_STRUC and Retrotecto©), referred to as the "All BoERVs" set or Dataset I, and (ii) using conserved BoERVs, that is, BoERVs whose *gag*, *pol* and *env* genes were detected by BLAST, which were referred to as the "Conserved BoERVs" set or Dataset II. Thus, it was possible to compare different types of BoERVs. Two additional datasets were also generated: (iii) BoERVs whose LTR and RT region were detected and (iv) BoERVs detected by at least two methods of the three methods used (BLAST, LTR_STRUC and Retrotecto©). The results obtained from the latter two datasets were redundant and are not shown here. These different datasets were successively more conservative to avoid any potential bias due to BoERV detection.

In each dataset, a search was performed to find genes in different positions and at different ranges around BoERVs. The search identified BoERVs included in genes, genes included in BoERVs and genes that were upstream or downstream from BoERVs by up to 5 and 50 kb. BoERVs included in genes were classified depending on their location in introns or exons.

To test if the detected associations could be explained by chance, a simulation was designed. BoERVs were placed in the genome randomly using a script written in PHP and 1000 simulations were built. Then the associations between BoERVs and genes were analyzed. Then, two expected distributions were built using a χ^2: 1) using the random values as observed and the average of the random values as expected; and 2) using the random values as observed and the value of a homogenous distribution as expected. Then the statistic of the observed associations was calculated using the average of 1) the random values and 2) the homogenous distribution values as expected values. This statistic was placed in those expected distribution to calculate the p-value. The comparison was carried out using R (R Development Core Team, 2008).

A χ^2 test was used to detect significant differences between the distribution of sense and antisense insertions of BoERVs as in Almeida et al. (2007). The orientation of BoERVs was assessed by comparing the direction of BoERVs to the direction of adjacent genes, with an expected value of 50% for each sense.

2.2. Functional characterization of genes near BoERVs

To identify overrepresented Gene Ontology terms in the genes from the datasets and ranges used, the FatiGO Compare tool of Babelomics (Al-Shahrour et al., 2006) was used. This tool was selected because it includes an annotation for cattle. The options used were as follows: gene list versus rest of genome; GO-biological process databases using default parameters; and a Fisher exact test. For all of the analyses performed, only terms using adjusted p values <0.05 were retrieved. The adjusted p values were calculated by FatiGO using the FDR procedure (Al-Shahrour et al., 2006).

2.3. Phylogenetic relationship of duplicated genes and related BoERVs

Genes annotated as "interferon" and located inside BoERVs were retrieved. Genes were aligned by MAFFT 6.833b (linsi option) (Katoh and Toh, 2008), and a phylogenetic analysis of nucleotide sequences was carried out using the neighbor-joining method implemented in MEGA 4 (Tamura et al., 2007). The pairwise deletion, 1000 bootstrap and K2P correction options were used. IFN-α (Accession numbers M10954, Z46504, AY325272, AY523531 and DQ396807), IFN-β (M15478), IFN-ω (M1102), IFN-τ (M31557, AF196320, AF196322, AF196324) and IFN-X (Walker and Roberts, 2009) bovine sequences were added as markers and used to classify selected sequences.

A phylogenetic analysis of BoERVs related to "IFN" genes was performed using the same procedure.

2.4. Detection of retroviral sequences in the bovine EST database

To detect expressed retroviral sequences, the "Conserved BoERV" dataset was used. The primary retroviral genes (*gag*, *pol* and *env*) were searched separately for from the sequences of the bovine EST database, which included a total of 1,517,053 ESTs, available on 28/08/2008 at the NCBI BLAST-server (http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=9913). The results were retrieved in XML format and were parsed and classified by in-house scripts written in PHP.

For the analyses, only matches with a bit-score value >200 and an e-value $<10^{-10}$ and for which the origin had been determined were considered (pooled ESTs were discarded). Those ESTs with positive matches with retroviral genes were classified according to tissue, organ and, if appropriate, developmental stage. In addition, the BoERV and retroviral gene matched pairs were annotated in each EST. If more than one BoERV matched a given EST, the EST was assigned to the BoERV with the longest matching sequence. GAPDH was used as a reference gene. To facilitate comparisons between tissues and organs and to avoid bias as a consequence of the number of ESTs in each tissue, the ratio, EST-R, as proposed by Stauffer et al. (2004), was used: (detected ESTs / total number of tissue ESTs) $* 10^5$. Statistical comparisons of retroviral genes against the reference gene were carried out using a test of equality implemented in R (R Development Core Team, 2008).

JASPAR database (Mathelier et al., 2013) was used to predict the transcription factor binding sites of selected BoERVs that are expressed in thyroids. The models related to "hormone-nuclear receptor" were selected and a "relative profile score threshold" of 95% was used.

2.5. Expression of genes related to expressed BoERVs

The search for genes found near BoERVs and expressed within the same tissue was performed in the same way as described above for BoERVs: sequences of genes located near the 81 expressed BoERVs were used for a BLAST search in the bovine EST database. Only ESTs with scores >200 and 100% identity were considered.

3. Results

3.1. BoERV/gene associations

To determine where bovine ERVs are located with respect to genes, the positions of bovine genes and bovine ERVs were compared. These analyses were carried out using two pools of ERVs: (i) all the BoERVs detected in our previous study (Garcia-Etxebarria and Jugo, 2010), referred to as the "All BoERVs" dataset or Dataset I, and (ii) a subset of Dataset I including the ERVs whose 3 main genes were detected by