Bovine leukemia virus can be classified into seven genotypes: evidence for the existence of two novel clades

Sabrina M. Rodriguez, Marcelo D. Golemba, Rodolfo H. Campos, Karina Trono and Leandro R. Jones

INTRODUCTION

Bovine leukemia virus (BLV) is a member of the family Retroviridae belonging to the genus Deltaretrovirus. This genus also includes *Simian T-lymphotropic virus 1, 2, 3 and 5* (STLV-1, -2, -3 and -5) and *Human T-lymphotropic virus 1, 2, 3 and 4* (HTLV-1, -2, -3 and -4). HTLV-3 and -4 have not yet been associated with any pathology, likely due to their recent identification and to the low number of infected cattle. Therefore, BLV is considered a model of HTLV-1 with any clinical signs (AL), and in approximately 30% of infected cattle, the virus causes a persistent lymphocytosis (PL) (Burny et al., 1987; Mirsky et al., 1996).

Analyses of the BLV envelope (env) gene of isolates collected in multiple geographical locations demonstrated significant sequence conservation (Camargos et al., 2002, 2007; Coulston et al., 1990). Nevertheless, up to seven BLV genotypes can be identified by RFLP analysis (Asfaw et al., 2005; Coulston et al., 1990; Fechner et al., 1997; Kettmann et al., 1981; Licursi et al., 2002). Furthermore, the env sequences of the BLV provirus from different locations worldwide have previously been classified into between two and four genetic groupings (Camargos et al., 2002, 2007; Felmer et al., 2005; Hemmatzadeh, 2007; Licursi et al., 2003; Mamoun et al., 1990; Monti et al., 2005; Zhao & Buehring, 2007), with similar results obtained from analyses of the *pol* gene (Asfaw et al., 1997) and the four complete genomes available (Dube et al., 2000). Some of these groupings appear to correlate with the geographical origin of the strains (Camargos et al., 2002; Coulston et al., 1990; Felmer et al., 2005; Hemmatzadeh, 2007; Mamoun et al., 1990; Monti et al., 2005; Zhao & Buehring, 2007). Each individual study gave unique names to the identified groups and, as of yet, no comprehensive analysis has integrated all of the available sequences. Thus, we have previously studies have classified the env sequences of bovine leukemia virus (BLV) provirus from different locations worldwide into between two and four genetic groupings. These different studies gave unique names to the identified groups and no study has yet integrated all the available sequences. Thus, we hypothesized that many of the different groups previously identified actually correspond to a limited group of genotypes that are unevenly distributed worldwide. To examine this hypothesis, we sequenced the env gene from 28 BLV field strains and compared these sequences to 46 env sequences that represent all the genetic groupings already identified. By using phylogenetic analyses, we recovered six clades, or genotypes, that we have called genotypes 1, 2, 3, 4, 5 and 6. Genotypes 1–5 have counterparts among the sequence groupings identified previously. One env sequence did not cluster with any of the others and was highly divergent when compared with the six genotypes identified here. Thus, an extra genotype, which we named 7, may exist. Similarity comparisons were highly congruent with phylogenetic analyses. Furthermore, our analyses confirmed the existence of geographical clusters.

The GenBank/EMBL/DDBJ accession numbers for the sequences identified in this study are FJ808571–FJ808598.

Five supplementary figures and a supplementary table are available with the online version of this paper.

The online version of this paper.

Previous studies have classified the env sequences of bovine leukemia virus (BLV) provirus from different locations worldwide into between two and four genetic groupings. These different studies gave unique names to the identified groups and no study has yet integrated all the available sequences. Thus, we hypothesized that many of the different groups previously identified actually correspond to a limited group of genotypes that are unevenly distributed worldwide. To examine this hypothesis, we sequenced the env gene from 28 BLV field strains and compared these sequences to 46 env sequences that represent all the genetic groupings already identified. By using phylogenetic analyses, we recovered six clades, or genotypes, that we have called genotypes 1, 2, 3, 4, 5 and 6. Genotypes 1–5 have counterparts among the sequence groupings identified previously. One env sequence did not cluster with any of the others and was highly divergent when compared with the six genotypes identified here. Thus, an extra genotype, which we named 7, may exist. Similarity comparisons were highly congruent with phylogenetic analyses. Furthermore, our analyses confirmed the existence of geographical clusters.

The GenBank/EMBL/DDBJ accession numbers for the sequences identified in this study are FJ808571–FJ808598.

Five supplementary figures and a supplementary table are available with the online version of this paper.
hypothesized that BLV could be a complex of several genotypes, such that many of the ‘different’ groups identified previously may actually correspond to a limited group of distinct genotypes unevenly distributed worldwide.

Herein, we investigated this hypothesis by studying the correspondence (and non-correspondence) between the previously identified sequence groupings in order to reveal precisely how many different phylogenetic clusters may actually exist. To do so, we combined new sequence data from 28 proviral DNAs generated in our laboratory with 46 env sequences representing all the groupings identified in previous studies and submitted the combined data to phylogenetic and similarity analyses.

METHODS

Provirus generation, PCR amplification and sequencing. The proviral DNAs corresponding to the 28 sequences reported here (Supplementary Table S1, available in JGV Online) were obtained from archived samples from a previously performed national BLV prevalence study (Trono et al., 2001). Total DNA from LSs was obtained following standard procedures (Sambrook et al., 1989) with some modifications as described below. Approximately 1 g tissue was homogenized in 5 volumes of extraction buffer (10 mM Tris/HCl pH 8, 0.5% SDS, 20 μg pancreatic RNase ml⁻¹) and clarified by centrifugation at 1519 g for 10 min at 4 °C, using a Sorvall SM24 rotor. Proteinase K was added to this homogenate to a final concentration of 100 μg ml⁻¹, and this mixture was incubated for 1 h at 42 °C. Following this incubation, 400 μl bi-distilled water was added to 100 μl of the preparation and nucleic acids were extracted by adding one volume of acid phenolchloroformisoamnicil (25:24:1), followed by precipitation with 0.1 volumes of 3 M sodium acetate (pH 5.2) and 2.5 volumes of ethanol. The DNA was pelleted by centrifugation for 10 min at maximum speed in a microcentrifuge and the pellets were washed with 70% ethanol, then dried and dissolved in 15 μl DNase/RNase-free water (Invitrogen).

Total DNA from blood samples from AL and PL animals was extracted using a commercial kit (REDExtract-N-Amp Blood PCR kit; Sigma) according to the manufacturer’s instructions. Briefly, 10 μl anticoagulated blood was incubated with 20 μl lysis solution at room temperature for 5 min, and the mixture was neutralized with 180 μl neutralization solution. Five microlitres of the neutralized extracts or of the LS sample DNA suspensions, was used as PCR templates.

A region of the env gene from positions 4833 to 6160 (reference sequence GenBank accession no. K02120) of the open reading frame was amplified by a hemi-nested PCR. The corresponding amino acid sequence encompasses residues 5–33 of the gp72 precursor signal peptide, the complete coding region of the gp51 surface (SU) glycoprotein and the first 214 aa of the gp30 transmembrane (TM) glycoprotein. The first amplification round was performed with primers 4583F (5′-TGGTTCCCTGGGTTT-3′) and 7514R (5′-AATCAAATGGGCAATTCT-3′). Then primers 4583F and 5997R (5′-CCTGATAAAATGCTTGACC-3′) were used in the second amplification round. AccuPrime Taq DNA polymerase (Invitrogen) was used for PCR amplification, following the manufacturer’s instructions. Thermal profiles for both rounds of amplification consisted of 30 cycles of 94 °C for 30 s, 55 °C for 30 s and 72 °C for 30 s. The cycles were preceded by an initial denaturation step at 94 °C for 30 s and followed by a final extension step at 72 °C for 5 min. Five microlitres of the first amplification round product was used as template for the second amplification round. The following controls were included in the amplification experiments: (i) the pBLV34 plasmid DNA (Van den Broeke et al., 1988), which contains the complete BLV genome (positive control), (ii) genomic DNA from a BLV-negative cow (first negative control) and (iii) a no template mix (second negative control). In parallel, the quality of the DNA used in all the amplification reactions was assessed by an endogenous control assay consisting of a PCR directed against subunit 8 of the bovine mitochondrial ATPase, as described elsewhere (Tartaglia et al., 1998). The env PCR amplicons were purified and sequenced directly using a primer walking strategy.

Construction and evaluation of the dataset. The 28 BLV env sequences from Argentina were combined with 46 BLV env sequences obtained from GenBank (Supplementary Table S1). Representatives of all the sequence groupings identified previously were included in the GenBank sequences selected (Camargos et al., 2002, 2007; Coulston et al., 1990; Dube et al., 2000; Hemmatzadeh, 2007; Mamoun et al., 1990; Molteni et al., 1996; Rice et al., 1984; Sagata et al., 1985; Willems et al., 1993; Zhao & Buehring, 2007). Prior to phylogenetic analysis, the sequences were aligned using the MAFFT program (Katoh et al., 2002, 2005) and the resulting alignment was inspected using the genetic data environment (GDE) program (Eisen, 1997; Smith et al., 1994) and further checked with the CLUSTAL_X program (Thompson et al., 2002). All the sequences were easily matched to each other and therefore the alignment did not contain any gap rich region, demonstrating that there were no difficulties in performing the alignment, which is the primary information used by phylogenetic programs (Phillips et al., 2000). This comparison indicated that the env gene is conserved enough to provide robust phylogenetic estimations, which is to say that phylogenetic trees should not depend on how the sequences are aligned, and that the addition of new data are not likely to result in large effects on the phylogenetic structure inferred here, other than the discovery of new groups that were not represented in our dataset.

Phylogenetic analyses. Four state-of-the-art applications implementating distance (neighbour-joining) (Saitou & Nei, 1987), Bayesian (Huelsenbeck et al., 2001; Rannala & Yang, 1996), maximum-likelihood (Felsenstein, 1981) and parsimony (Camin & Sokal, 1965; Farris, 1983) algorithms were used for phylogenetic analysis. For the model-based methods, neighbour-joining, maximum-likelihood and Bayesian, a DNA substitution model was obtained using the MrAIC script (Nylander, 2004). This analysis indicated that the model with the best fit for analysis of the BLV env sequences, with the smallest number of parameters, was the model described by Hasegawa et al. (1985), which accounts for different transition and transversion rates and unequal base composition, including site rate variation modelled by a gamma distribution (HKY+G), which indicates that substitution rates along different positions of the gene are heterogeneous.

The maximum-likelihood inference was performed with the PHYML program (Guindon & Gascuel, 2003). The advantage of PHYML over traditional programs is that the implemented algorithms (Guindon & Gascuel, 2003) allow for the analysis of large datasets (>40–50 sequences) that, due to the underlying computational complexity of maximum-likelihood methods, are limiting for traditional approaches (Chor et al., 2000; Guindon & Gascuel, 2003; Swofford et al., 1996), a difficulty that worsens if resampling analyses are needed. We allowed PHYML to estimate all the model parameters during tree searches.

The Bayesian analyses were performed with the MrBayes program (Huelsenbeck & Ronquist, 2001; Ronquist & Huelsenbeck, 2003). There are three critical aspects of Bayesian analyses: (i) to ensure that the Markov chain reaches convergence or stationarity, (ii) to achieve good mixing during the sampling process and (iii) to ensure that the analyses do not get trapped in local optima. All sample points prior to
reaching stationarity are essentially random and should be discarded as burn-in samples as they are not considered to contain useful information. Our analyses were run independently twice, and Metropolis-coupled Markov Chain Monte Carlo (MCMC) was used to enhance the tree-climbing capabilities of the Markov chains (Huelsenbeck & Ronquist, 2001). We used eight incrementally heated Markov chains following the default heating function of MrBayes. Every tenth generation, ten attempts were made to swap states between pairs of chains picked at random. MrBayes was run in a computer cluster using four processors per run. We first ran 10 × 10^6 generations of the MCMC process using a sample rate of 100. The plotting of the log-likelihood scores against generation times indicated that the two independent runs, which started from random trees, reached stationarity around generation 50000 and at equivalent average log likelihoods (first run, mean = −4869.942, median = −4869.521; second run, mean = −4869.993, median = −4869.281). The final analyses were run for 2 × 10^7 generations with a sampling frequency of 1000, thus obtaining two samples of 20000 topologies from the posterior distribution of trees. The MCMC processes converged at the same log-likelihoods (first run: mean = −4869.941, median = −4869.592; second run: mean = −4869.701, median = −4869.384). The effective sample sizes of each estimated parameter, which were calculated with the Tracer program (Rambaut & Drummond, 2007), were greater than 300 and the standard deviation of splits frequencies was <0.004. After discarding the first 100 trees as burn-in, the combined tree samples were summarized by generating a 50% majority rule consensus tree. The clades’ posterior probabilities were estimated by the percentage of as burn-in samples as they are not considered to contain useful information. Our analyses were run independently twice, and Metropolis-coupled Markov Chain Monte Carlo (MCMC) was used to enhance the tree-climbing capabilities of the Markov chains (Huelsenbeck & Ronquist, 2001). We used eight incrementally heated Markov chains following the default heating function of MrBayes. Every tenth generation, ten attempts were made to swap states between pairs of chains picked at random. MrBayes was run in a computer cluster using four processors per run. We first ran 10 × 10^6 generations of the MCMC process using a sample rate of 100. The plotting of the log-likelihood scores against generation times indicated that the two independent runs, which started from random trees, reached stationarity around generation 50000 and at equivalent average log likelihoods (first run, mean = −4869.942, median = −4869.521; second run, mean = −4869.993, median = −4869.281). The final analyses were run for 2 × 10^7 generations with a sampling frequency of 1000, thus obtaining two samples of 20000 topologies from the posterior distribution of trees. The MCMC processes converged at the same log-likelihoods (first run: mean = −4869.941, median = −4869.592; second run: mean = −4869.701, median = −4869.384). The effective sample sizes of each estimated parameter, which were calculated with the Tracer program (Rambaut & Drummond, 2007), were greater than 300 and the standard deviation of splits frequencies was <0.004. After discarding the first 100 trees as burn-in, the combined tree samples were summarized by generating a 50% majority rule consensus tree. The clades’ posterior probabilities were estimated by the percentage of samples recovering any particular clade.

The neighbour-joining trees were obtained with the PAUP* program (Swofford, 1998) with the evolutionary model set to HKY + G. The mean transition/transversion rate (mean = 8.67, median = 8.55) and shape parameter of the gamma distribution (mean = 0.080, median = 0.081) were obtained from the posterior sample of parameters from the Bayesian analyses and the empirical base frequencies were used. The Parsimony analyses were performed with TNT (Giribet, 2005; Goloboff et al., 2008), which is a relatively new program implementing novel technologies of phylogenetic analyses that aims to provide a thorough exploration of the tree space (Goloboff, 1999; Goloboff & Farris, 2001; Nixon, 1999), ensuring that all the possible phylogenetic hypotheses that could be supported by the data are considered. In light of the inequality between transition and transversion rates and the existence of among-sites rate heterogeneity detected by the probabilistic analyses, we decided to use a weighted character substitution matrix and to model the rate heterogeneity. Based on the estimates obtained from the Bayesian posterior sample of parameters, we first set the transition and transversion costs to 1 and 8, respectively. The existence of different rates among sites implies that some sites may be more homoplasious than others, indicating that the characters’ fit to the trees is expected to vary from one site to another. Therefore, we further modelled the among-site rate heterogeneity by using the implied weights strategy described by Goloboff (1993). This strategy consists of weighting characters non-iteratively and is based on a concave function of homoplasy. This results in a less influential impact of differences in steps occurring on characters that show more homoplasy on the trees. For tree searches, we built 1000 Wagner trees (Farris, 1970) by random addition sequence (RAS) of terminals (Goloboff, 1999) and the resulting topologies were submitted to tree bisection reconnection (TBR) branch swapping (Swofford et al., 1996). One hundred trees were held while swapping and ambiguously supported branches were automatically collapsed during tree searches. This analysis resulted in a fully stable consensus tree (i.e. a tree that no longer changed by the addition of further RAS + TBR cycles), indicating that every possible consensus topology that could be supported by the data were represented among the found trees (Goloboff, 1999; Goloboff & Farris, 2001).

Bootstrap analyses (Felsenstein, 1985) were applied to evaluate the robustness of the neighbour-joining, parsimony and maximum-likelihood topologies using PAUP*, TNT and PHYLML, respectively.

**Similarity analyses.** The number of nucleotide or amino acid substitutions (D) between a given pair of sequences, was obtained by using the following formula:

\[
D_{ab} = \sum_{i=1}^{n} f(a_i, b_i)
\]

P is calculated as the number of positions in the alignment of the sequences a and b, and f(a_i, b_i) is calculated as follows:

\[
f(a_i, b_i) = \begin{cases} 
0, & \text{if } a_i = b_i \\
1, & \text{if } a_i \neq b_i
\end{cases}
\]

The inter- and intra-genotype comparisons were averaged by the number of sequence comparisons performed in each case. For example, there were 37 sequences in genotype 1; thus, the number of pair-wise comparisons inside genotype 1 is given by a combination of 37 elements taken in pairs:

\[
\frac{n!}{2!(n-k)!} = \frac{37!}{2(37-2)!} = \frac{37!}{2(35)!} = 666
\]

Likewise, the mean number of pair-wise differences between two genotypes, x and y, with n_x and n_y sequences each, was obtained by the following formula:

\[
\Delta_{xy} = \frac{2(n_x + n_y - 2)!}{(n_x + n_y)!} \prod_{1 \leq i < j \leq n_x}^{n_y} f(a_i, b_j)
\]

These algorithms were implemented in a script written in the R statistical package language (http://www.R-project.org), which is available from the authors upon request.

**RESULTS**

A dataset of 74 env sequences generated from multiple BLV proviruses from different geographical locations (Supplementary Table S1) was analysed by phylogenetic techniques and comparisons of sequence similarities. The aligned sequences contained 1329 nt, of which 262 were variable and 142 were phylogenetically informative. The informative variability was homogeneously distributed along the dataset (Supplementary Fig. S1). The mean rate of non-synonymous to synonymous substitutions (D_{s0}/D_{s0}) was 0.331, indicating that most of the observed variability is neutral (Supplementary Fig. S2). Furthermore, synonymous substitutions were homogeneously distributed along the sequences (Supplementary Fig. S2). Interestingly, the majority of amino acid substitutions in the SU glycoprotein were located at known antigenic determinants (Fig. 1) (Ban et al., 1992; Bruck et al., 1982a, b, 1984a, b; Callebaut et al., 1991, 1993; Gatei et al., 1993; Portetelle et al., 1989). This uneven distribution of non-synonymous variability strongly suggests that mutation at these positions of the env gene could be driven by the host immune response.
All the phylogenetic methods indentified six sequence clusters, which we called genotypes 1, 2, 3, 4, 5 and 6 (Fig. 2, Supplementary Figs S3–S5, Table 1 and Supplementary Table S1). These groupings were supported by moderate to high bootstrap values and by high posterior probabilities (Table 1, Fig. 2, Supplementary Figs S3–S5). Genotypes 2, 3, 4 and 6 displayed bootstrap values above 98, whereas the supports for genotype 5 were 72 with the maximum-likelihood method and 100 with the parsimony and neighbour-joining approaches. Genotype 1 had a support of 89 in the parsimony analysis (Fig. 2). This group was supported by values of 95 in the maximum-likelihood analyses and 97 in the neighbour-joining analysis (Table 1, Supplementary Figs S3 and S5). A strain from Italy (S83530), could not be assigned to any of these six genotypes (Fig. 2, Supplementary Figs S3–S5). Camargos et al. (2007) observed that, in some of their analyses, S83530 clustered with strains from Chile (AF515280, AY515276 and AY515274) and Brazil (AY185360), albeit with a low bootstrap value (their cluster 4). The clustering of S83530, AF515280, AY515276 and AY515274 was also observed by Felmer et al. (2005) (their group IV). The sequences AF515280, AY515276 and AY515274 are partial env sequences of 444 bases and were therefore not included in our dataset in anticipation that such large amounts of missing characters may have resulted in unresolved or incorrect trees and artefactual branch supports in our analyses (McMahon & Sanderson, 2006; Smith & Donoghue, 2008; Wiens, 2006). Our analyses clearly showed that AY185360 does not cluster with S83530 but with the Argentinean sequence FJ808582 (genotype 6; Fig. 2, Supplementary Figs S3–S5, Table 1). The inability of S83530 to cluster in any of the six genotypes identified here indicates that this strain could belong to a seventh genetic group, which we called genotype 7.

Genotypes 1–5 have one or more counterparts among the groupings identified previously, indicating that the different groups characterized in past studies correspond, in many cases, to the same genotype (Table 1). The sequences studied by Zhao & Buehring (2007), who identified four clusters of sequences – US Californian, Consensus, European and Costa Rican – were all included in our dataset. The strains from the Consensus cluster grouped to our genotype 1, whereas the European cluster was represented in our genotype 4. The clusters US

Fig. 1. Distribution of amino acid substitutions along the surface glycoprotein gp51. Variable amino acid positions of the env gene product as well as the corresponding polymorphisms are indicated. ND1, ND2 and ND3: neutralizing domains 1, 2 and 3, respectively (Callebaut et al., 1993). A, B, D and E indicate viral epitopes described by Bruck et al. (1982a, b), Portetelle et al. (1989) and Ban et al. (1992). F–H indicate antigenic determinants identified by Bruck et al. (1982a, b), Portetelle et al. (1989) and Callebaut et al. (1991). CD4+, CD8+ indicate CD4+ and CD8+ T-cell epitopes (Gatei et al., 1993).
Californian and Costa Rican correspond to our genotypes 3 and 5, respectively. The sequences M35240, AF257515.1 and D00647.1 were included in clusters I, II and III described by Felmer et al. (2005), respectively. This indicates that clusters I, II and III identified by this group are equivalent to our genotypes 4, 2 and 1, respectively (Table 1). Clusters 1, 2 and 3 described by Camargos et al. (2007) included the sequences AF257515.1, AF503581.1 and M35242.1, respectively, which, in our trees, were clustered into genotypes 2, 1 and 4. Thus, our genotypes 1, 2 and 4 are equivalent to clusters 2, 1 and 3 described by Camargos et al. (2007) (Table 1). The sequences AF257515.1, K02120.1 and M35238.1, which clustered into genotypes 2, 1 and 4, were included in the first, third and second clusters described by Monti et al. (2005) (Table 1). The two subgroups identified by Mamoun et al. (1990),
Japan/USA and Belgium/France, correspond to our genotypes 1 and 4, and were represented in our dataset by the sequences M35242.1 and M35238.1, respectively (Table 1). Our trees were also in agreement with the clustering scheme of Licursi et al. (2003), who divided BLV stains into Japanese, Argentinean and European groups. The Japanese cluster, represented by strains D00647.1, M35242.1 and M35239.1 in our dataset, corresponded to our genotype 1 (Table 1), whereas our genotype 2 corresponds to their Argentinean cluster. Licursi et al. (2003) included the Italian sequence (S83530) in the European group, which also included M35238.1 and M35240.1 in our study. Therefore, with the exception of S83530, our genotype 4 corresponds to the European group of Licursi et al. (2003). The sequences from Iran clustered into a single branch of genotype 1 and were more related to the Australian sequence analysed, in concordance with the analyses of Hemmatzadeh (2007). Based on RFLP analyses of env sequences, Fechner et al. (1997) classified BLV isolates into seven groups (A–G). Their sequence comparisons indicated that strains from groups A and D were genetically similar to the env sequence of the LB59 strain (M35238.1). They also observed that the sequences from groups B and G were similar to the env sequences from the Australian, BLV-Jap and FLK-BLV strains (D00647.1, K02120.1 and M35242, respectively). Thus, the RFLP groups A and D correspond to our genotype 4, whereas the B and G groups would correspond to our genotype 1.

In order to compare the results of phylogenetic analyses with the similarity relationships derived from sequence comparisons, we compared each possible pair of strains by counting the number of substitutions at the nucleotide and amino acid levels. The outcomes of these analyses were highly congruent with the phylogenetic trees. At the nucleotide level, the mean number of pair-wise nt substitutions between sequences from the same genotype was 10.6, with a minimum of 6.5 and a maximum of 19 (Table 2). Conversely, the mean number of pair-wise nt substitutions between sequences from different genotypes was 40, with a minimum value of 34 and a maximum value of 48.6 (Table 2). Furthermore, the mean number of aa substitutions at the intra-genotype level was 3.3 (minimum

Table 1. BLV genotypes identified, statistical support and their counterparts in previous studies

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Support</th>
<th>Counterparts</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1/100/100/100</td>
<td>US Californian, Zhao &amp; Buehring (2007)</td>
</tr>
<tr>
<td>5</td>
<td>1/100/72/100</td>
<td>Costa Rican cluster, Zhao &amp; Buehring (2007)</td>
</tr>
<tr>
<td>6</td>
<td>1/100/100/100</td>
<td>–</td>
</tr>
</tbody>
</table>

*Genotypes identified here through phylogenetic analysis (Fig. 1). †Posterior probabilities and bootstrap supports obtained in this work: posterior probability/parsimony/maximum-likelihood/neighbour-joining. ‡Groupings homologous to the indicated genotype. –, No counterparts in previous studies.

Table 2. Mean number of nucleotide substitutions between all sequence pairs from each genotype (intra-genotype) or between sequences from a given genotype against all the sequences from other genotypes (inter-genotype)

<table>
<thead>
<tr>
<th>Genotype</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-genotype</td>
<td>9.0</td>
<td>7.1</td>
<td>6.5</td>
<td>14.0</td>
<td>19.0</td>
<td>8.5</td>
<td>*</td>
</tr>
<tr>
<td>Inter-genotype</td>
<td>38.4</td>
<td>38.5</td>
<td>37.5</td>
<td>40.6</td>
<td>48.6</td>
<td>42.8</td>
<td>34</td>
</tr>
</tbody>
</table>

*No intra-genotype substitutions are seen as genotype 7 comprises only one strain.
Our analyses confirmed that, as observed previously, phylogenetic and similarity groupings correlate with the geographical origin of the provirus (Supplementary Table S1; \( P=6.838 \times 10^{-15}, \chi^2 \) test; \( P=4.583 \times 10^{-12}, \) Fisher’s exact test). The Belgian and French sequences (\( n=5 \)) all clustered to genotype 4 and the sequences isolated from Japan and the USA were clustered exclusively in genotypes 1 (17/21) and 3 (3/21), respectively. Likewise, Iranian strains clustered within a single branch nested into genotype 1 (Fig. 2, Supplementary Figs S3–S5). The most obvious correspondences between genotype and geographical origin were observed for genotypes 2 and 5, in which genotype 2 was composed almost exclusively (15/16) of sequences from BLV strains from Argentina, and genotype 5, in agreement with the work by Zhao & Buehring (2007), included sequences exclusively from Costa Rica (Fig. 2, Supplementary Figs S3–S5).

**DISCUSSION**

The analyses described here show that BLV strains can be classified into seven genotypes (Fig. 2, Supplementary Figs S3–S5, Tables 1, 2 and 3). Genotypes 1–5 have one or more counterparts among 20 groupings identified previously (Table 1). The genotype and geographical origin of each particular BLV strain were highly correlated (Supplementary Table S1; \( P=6.838 \times 10^{-15}, \chi^2 \) test; \( P=4.583 \times 10^{-12}, \) Fisher’s exact test). These results demonstrate that the different BLV genotypes have an uneven geographical distribution.

A question that naturally arises is why a virus infecting a single host is diversified into discrete genetic clusters. Others have identified a range of factors that could account for genetic groupings for other viruses. In the picornavirus, for example, structural genes cluster according to viral serotypes (Simmonds, 2006). Similarly, in psittacid herpesviruses, the serotypes of the virus isolates could be predicted by their genotypes (Tomaszewski et al., 2003). For bovine pestivirus, it has been shown that strains belonging to different genotypes lead to altered clinical manifestations in infected cattle (Baule et al., 1997, 2001; Jones et al., 2001), which resulted in the proposal that the diversification of this virus could be driven by viral adaptation in order to use multiple different pathogenic strategies (Jones et al., 2004). Neither of these alternatives explains BLV genetic structuring, as the virus cannot be divided into unique serotypes and there is no apparent relation between the genotype of the infecting viral strains and the clinical manifestations of BLV disease. Other viruses have diversified into discrete genetic clusters as a result of the hosts’ dispersal, possibly due to founder effects and genetic drift. For instance, there is a strong dependence between the geographical origin and the genotype of Kaposi’s sarcoma-associated herpesvirus (KSHV) isolates that has been attributed to the expansion of the distinct KSHV-infected populations following the major human migrations out of Africa and to the rest of the world (Hayward & Zong, 2007; Zong et al., 1999). Likewise, correlations between genetic clustering and the geographical origin of particular viruses have been demonstrated for STLV (Makuwa et al., 2004) and HTLV (Cassar et al., 2007; Dube et al., 1993; Eirin et al., 2008; Vidal et al., 1994). For these viruses, it is also believed that this correlation is due to the migratory flow of their hosts (Cassar et al., 2007; Dube et al., 1993; Makuwa et al., 2004; Vidal et al., 1994). Therefore, a very plausible explanation that may account for the radiation of BLV genotypes could be that the diversification of the virus has been driven by the historical dispersion of its host, as the worldwide cattle population is closely linked to human colonization, ethnic history and animal domestication over the last 200–1000 years. Spread of the virus throughout animal populations associated with human migration implies a leptokurtic form of dispersion in which long-distance spread is achieved by small proportions of the source population (Ibrahim et al., 1996). When dispersal proceeds in this manner, geographically isolated groups are established within which the genetic variation is lower than between these unique groups and individuals from other, distinct geographical regions. This theory is supported by the analyses presented here, as our trees were characterized by the presence of clusters of relatively similar sequences separated from each other by relatively longer branches (Fig. 2, Supplementary Figs S3–S5, Tables 2 and 3). Furthermore, the occurrence of long-distance dispersal achieved by a minority fraction of the virus population also supports the concept of viral transmission via close contact among individuals, which has been argued as an explanation for the characteristics of viral dispersion and endemic patterns in the related virus, HTLV (Cassar et al., 2007; Dube et al., 1993; Eirin et al., 2008; Vidal et al., 1994). As suggested previously (Camargos et al., 2007; Zhao & Buehring, 2007), the late strains dispersal caused by modern cattle trading may account for the current presence of more than a single group of viruses in certain geographical areas.

Although the molecular variation observed here affected primarily the nucleotide sequences, there were significant

### Table 3. Mean number of amino acid substitutions between all sequence pairs from each genotype (intra-genotype) or between sequences from a given genotype against all the sequences from other genotypes (inter-genotype)

<table>
<thead>
<tr>
<th>Genotype:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-genotype</td>
<td>3.6</td>
<td>3.0</td>
<td>0.4</td>
<td>3.6</td>
<td>6.2</td>
<td>3</td>
<td>*</td>
</tr>
<tr>
<td>Inter-genotype</td>
<td>9.9</td>
<td>8.5</td>
<td>7.9</td>
<td>10.2</td>
<td>13.6</td>
<td>10.3</td>
<td>8.6</td>
</tr>
</tbody>
</table>

*No intra-genotype substitutions are seen as genotype 7 comprises only one strain.*
numbers of amino acid substitutions at both the intra- and inter-genotype levels (Table 3). Furthermore, an important proportion of these amino acid substitutions were concentrated within known epitopes of the gp51 protein (Fig. 1), in agreement with previous molecular evolutionary analyses (Zhao & Buehring, 2007). As gp51 is a target for BLV detection assays, these observations raise concerns about whether BLV variability could affect the ability to identify viruses in all infected populations. Sequence variability within unique viral isolates may result in antigenic differences that could be linked to the failure of molecular and serological detection methods. In fact, it has been shown that the sensitivity and specificity of different serological testing methods already demonstrate significant variation (Trono et al., 2001) and produce different results when compared with molecular methods (Fechner et al., 1996; Reichel et al., 1998). Furthermore, the existence of BLV strains that escape antibody detection, associated with the presence of particular genotypes, has been described (Fechner et al., 1997; Monti et al., 2005). Monti et al. (2005) observed that 31 of 445 animals that tested positive by PCR were negative when screened by serological analysis. Moreover, sequence analyses of the env genes from provirus isolated from two of these cows indicated that they belonged to a divergent genotype. Likewise, Fechner et al. (1997) observed a link between the genotype of particular proviruses and the failure of commonly used detection methods. It is possible that the disagreement between different diagnostic techniques may be explained, at least in part, by the existence of multiple genetic backgrounds among the virus samples examined. Likewise, the variability of BLV may also be important for vaccine development, especially in transgenic approaches, for which gp51 is the immunogen of choice (Brillowska et al., 1999; Kerkhofs et al., 2000).

ACKNOWLEDGEMENTS

Continuous support from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) is greatly appreciated. The pBLV344 plasmid was kindly provided by Dr Luc Willem, Department of Applied Biochemistry and Biology, Faculty of Agronomy, B5030 Gembloux, Belgium. During part of this work, L. R. J. was a Research Scholar at the department of Microbiology, New England Primate Research Center (Harvard Medical School). We are in debt to Melissa Laird for manuscript proofreading. We thank Luc Willem and two anonymous reviewers for their useful comments and suggestions. During part of this work S. M. R. was a FONCyT (Agencia Nacional de Promociones Científicas) fellow.

REFERENCES


Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2002). Multiple sequence alignment using *CLUSTAL_W* and *CLUSTAL_X*. *Carr Protec Bioinformatics* 2, Unit 2.3.


BLV can be classified into seven genotypes