Journal of General Virology (2009), **90**, 2788–2797 **DOI** 10.1099/vir.0.011791-0

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

Sabrina M. Rodriguez,[1] Marcelo D. Golemba,[2] Rodolfo H. Campos,[2] Karina Trono[1] and Leandro R. Jones[3]

Correspondence
Leandro R. Jones
lrj@efpu.org.ar

[1]Instituto de Virología, CNIA, INTA-Castelar, Argentina

[2]Cátedra de Virología, Facultad de Farmacia y Bioquímica, UBA, Argentina

[3]Division of Molecular Biology, Estación de Fotobiología Playa Unión, CC 15, Rawson, Chubut 9103, Argentina

Received 18 April 2009
Accepted 6 July 2009

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.

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 isolates. Therefore, BLV is considered a model of HTLV-1 and -2 (Willems *et al.*, 2000). BLV is recognized as the aetiological agent of enzootic bovine leukosis, a disease that results in significant economic losses for the worldwide cattle industry. The most conspicuous clinical manifestation of bovine leukosis, which only develops in a small fraction of infected animals, is the clonal expansion and local accumulation of B cells that results in the development of lymphoid tumours (lymphosarcoma, LS) (Gillet *et al.*, 2007). The majority of infections are not associated 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 (Dube *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

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 phenol:chloroform:isoamilic (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′-TGGGTTCCCTGGCGTTT-3′) and 7514R (5′-AATCAAATGGCCAATTTT-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 3 min 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 pBLV344 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 implementing 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 50 000 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 20 000 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 PHYML, 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:

$$D_{ab} = \sum_{i=1}^{P} f(a_i, b_i)$$

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:

$$\frac{n!}{2!(n-k)!} = \frac{37!}{2(37-2)!} = 666$$

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:

$$\Delta_{x,y} = \frac{2[(n_x + n_y) - 2]! \sum_{a_x=1}^{a_x=n_x} \sum_{b_y=1}^{b_y=n_y} D_{ab}}{(n_x + n_y)!}$$

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_N/d_S) 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.

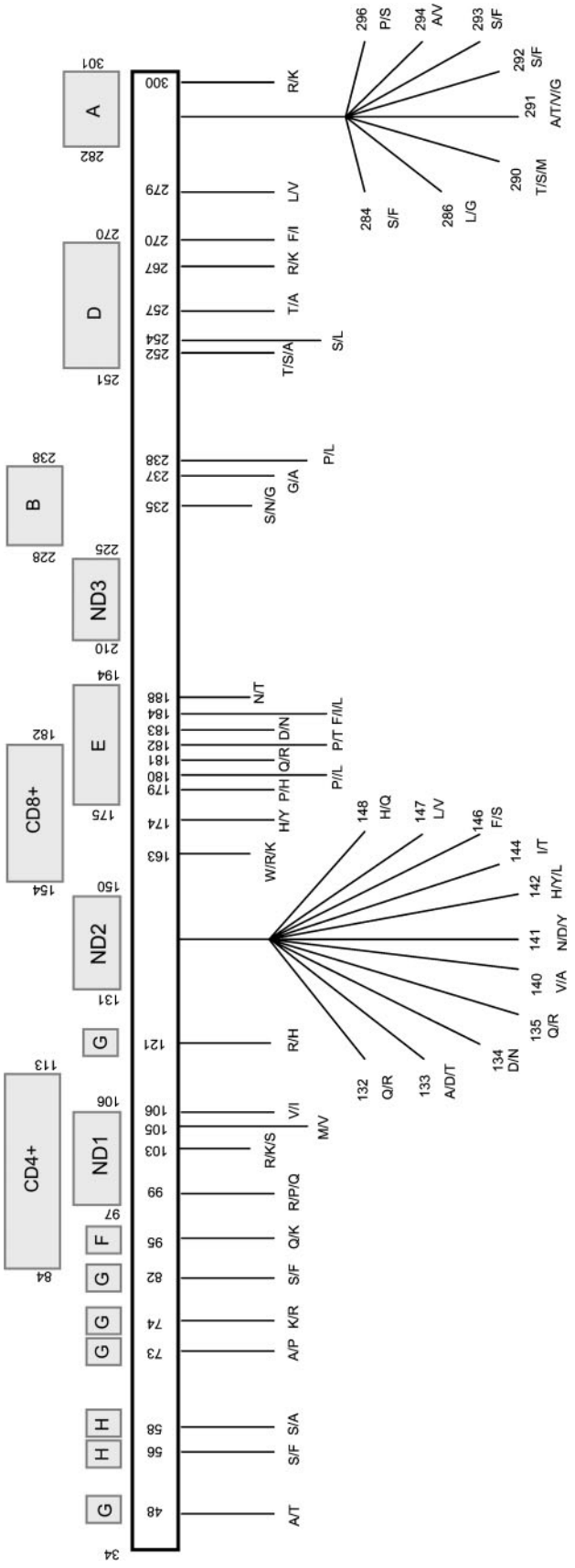

Fig. 1. Distribution of amino acid substitutions along the surface glycoprotein gp51. Variable amino acid positions of the *env* gene product as well as the corresponding polymorphisms are indicated. ND1, ND2 and ND3: neutralizing domains 1, 2 and 3, respectively (Callebaut *et al.*, 1993). A, B, D and 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).

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

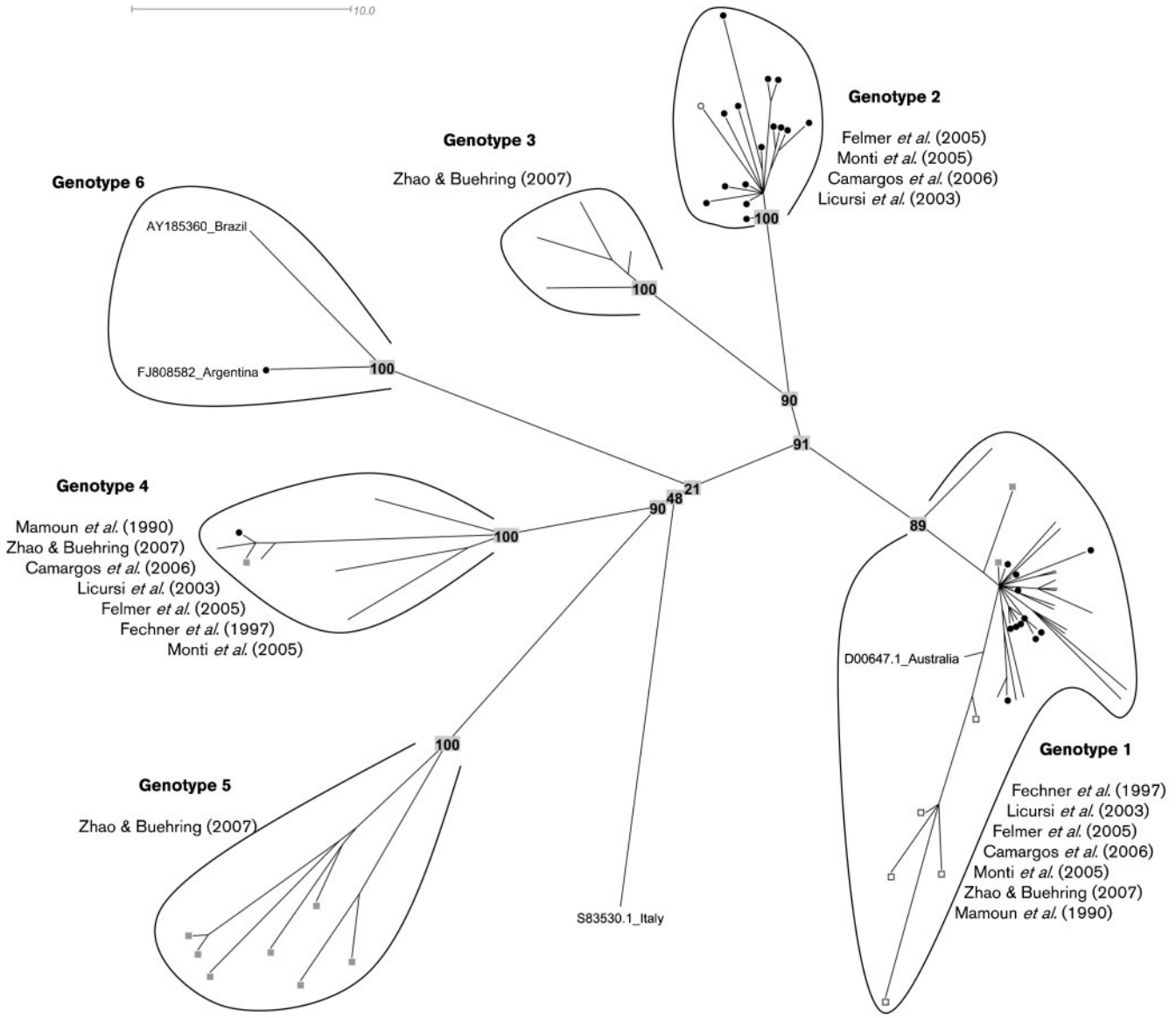

Fig. 2. Phylogenetic tree of 74 *env* sequences from different geographical locations worldwide (Supplementary Table S1). Genotypes 1–6 identified here are indicated with lines surrounding the corresponding tree branches. The citations given next to the genotypes indicate previous studies that support the existence of those genotypes (see Table 1 for details). The tree is a strict consensus tree obtained by the TNT program. Numbers on internal branches indicate bootstrap supports. Branch lengths are proportional to the number of nucleotide substitutions (bar, 10 substitutions). We obtained equivalent topologies with the Bayesian, neighbour-joining and maximum-likelihood phylogenetic methods (Table 2, Supplementary Figs S3–S5). Sequences originating from Argentina (●), Costa Rica (■), from the locality of Minas Gerais, which is located in the south of Brazil, (○) (in genotype 2) and Iran (□) are indicated.

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),

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

Genotype*	Support†	Counterparts‡
1	1/89/95/97	Japan/USA subgroup, Mamoun *et al.* (1990) Groups B and G, Fechner *et al.* (1997) Japanese group, Licursi *et al.* (2003) Group III, Felmer *et al.* (2005) Third cluster, Monti *et al.* (2005) Cluster 2, Camargos *et al.* (2007) Consensus cluster, Zhao & Buehring (2007)
2	1/100/100/100	Group II, Felmer *et al.* (2005) Argentinean group, Licursi *et al.* (2003) First cluster, Monti *et al.* (2005) Cluster 1, Camargos *et al.* (2007)
3	1/100/100/100	US Californian, Zhao & Buehring (2007)
4	1/100/99/98	Belgium/France subgroup, Mamoun *et al.* (1990) Groups A and D, Fechner *et al.* (1997) European cluster, Zhao & Buehring (2007) Second cluster, Monti *et al.* (2005) Cluster 3, Camargos *et al.* (2007) Group I, Felmer *et al.* (2005)
5	1/100/72/100	Costa Rican cluster, Zhao & Buehring (2007)
6	1/100/100/100	–

*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.

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.

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)

Genotype:	1	2	3	4	5	6	7
Intra-genotype	9.0	7.1	6.5	14.0	19.0	8.5	*
Inter-genotype	38.4	38.5	37.5	40.6	48.6	42.8	34

*No intra-genotype substitutions are seen as genotype 7 comprises only one strain.

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

0.4, maximum 6.2), whereas it was 9.8 (minimum 7.9, maximum 13.6) at the inter-genotype level.

Our analyses confirmed that, as observed previously, phylogenetic and similarity groupings correlate with the geographical origin of the provirus (Supplementary Table S1; $P=6.838\times10^{-15}$, χ^2 test; $P=4.583\times10^{-12}$, 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; $P=6.838\times10^{-15}$, χ^2 test; $P=4.583\times10^{-12}$, 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)

Genotype:	1	2	3	4	5	6	7
Intra-genotype	3.6	3.0	0.4	3.6	6.2	3	*
Inter-genotype	9.9	8.5	7.9	10.2	13.6	10.3	8.6

*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 Willems, 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 Willems 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

Asfaw, Y., Tsuduku, S., Konishi, M., Murakami, K., Tsuboi, T., Wu, D. & Sentsui, H. (2005). Distribution and superinfection of bovine leukemia virus genotypes in Japan. *Arch Virol* **150**, 493–505.

Ban, J., Czene, S., Altaner, C., Callebaut, I., Krchnak, V., Merza, M., Burny, A., Kettmann, R. & Portetelle, D. (1992). Mapping of sequential epitopes recognized by monoclonal antibodies on the bovine leukaemia virus external glycoproteins expressed in *Escherichia coli* by means of antipeptide antibodies. *J Gen Virol* **73**, 2457–2461.

Baule, C., van Vuuren, M., Lowings, J. P. & Belak, S. (1997). Genetic heterogeneity of bovine viral diarrhoea viruses isolated in Southern Africa. *Virus Res* **52**, 205–220.

Baule, C., Kulcsar, G., Belak, K., Albert, M., Mittelholzer, C., Soos, T., Kucsera, L. & Belak, S. (2001). Pathogenesis of primary respiratory disease induced by isolates from a new genetic cluster of bovine viral diarrhea virus type I. *J Clin Microbiol* **39**, 146–153.

Brillowska, A., Dabrowski, S., Rulka, J., Kubis, P., Buzala, E. & Kur, J. (1999). Protection of cattle against bovine leukemia virus (BLV) infection could be attained by DNA vaccination. *Acta Biochim Pol* **46**, 971–976.

Bruck, C., Mathot, S., Portetelle, D., Berte, C., Franssen, J. D., Herion, P. & Burny, A. (1982a). Monoclonal antibodies define eight independent antigenic regions on the bovine leukemia virus (BLV) envelope glycoprotein gp51. *Virology* **122**, 342–352.

Bruck, C., Portetelle, D., Burny, A. & Zavada, J. (1982b). Topographical analysis by monoclonal antibodies of BLV-gp51 epitopes involved in viral functions. *Virology* **122**, 353–362.

Bruck, C., Portetelle, D., Mammerickx, M., Mathot, S. & Burny, A. (1984a). Epitopes of bovine leukemia virus glycoprotein gp51 recognized by sera of infected cattle and sheep. *Leuk Res* **8**, 315–321.

Bruck, C., Rensonnet, N., Portetelle, D., Cleuter, Y., Mammerickx, M., Burny, A., Mamoun, R., Guillemain, B., van der Maaten, M. J. & Ghysdael, J. (1984b). Biologically active epitopes of bovine leukemia virus glycoprotein gp51: their dependence on protein glycosylation and genetic variability. *Virology* **136**, 20–31.

Burny, A., Cleuter, Y., Kettmann, R., Mammerickx, M., Marbaix, G., Portetelle, D., Van den Broeke, A., Willems, L. & Thomas, R. (1987). Bovine leukaemia: facts and hypotheses derived from the study of an infectious cancer. *Cancer Surv* **6**, 139–159.

Callebaut, I., Burny, A., Krchnak, V., Gras-Masse, H., Wathelet, B. & Portetelle, D. (1991). Use of synthetic peptides to map sequential epitopes recognized by monoclonal antibodies on the bovine leukemia virus external glycoprotein. *Virology* **185**, 48–55.

Callebaut, I., Voneche, V., Mager, A., Fumiere, O., Krchnak, V., Merza, M., Zavada, J., Mammerickx, M., Burny, A. & Portetelle, D. (1993). Mapping of B-neutralizing and T-helper cell epitopes on the bovine leukemia virus external glycoprotein gp51. *J Virol* **67**, 5321–5327.

Camargos, M. F., Stancek, D., Rocha, M. A., Lessa, L. M., Reis, J. K. & Leite, R. C. (2002). Partial sequencing of *env* gene of bovine leukaemia virus from Brazilian samples and phylogenetic analysis. *J Vet Med B Infect Dis Vet Public Health* **49**, 325–331.

Camargos, M. F., Pereda, A., Stancek, D., Rocha, M. A., dos Reis, J. K., Greiser-Wilke, I. & Leite, R. C. (2007). Molecular characterization of the *env* gene from Brazilian field isolates of bovine leukemia virus. *Virus Genes* **34**, 343–350.

Camin, J. H. & Sokal, R. R. (1965). A method for deducing branching sequences in phylogeny. *Evolution* **19**, 311–326.

Cassar, O., Capuano, C., Bassot, S., Charavay, F., Duprez, R., Afonso, P. V., Abel, M., Walter, H., Mera, W. & other authors (2007). Human T lymphotropic virus type 1 subtype C melanesian genetic variants of the Vanuatu Archipelago and Solomon Islands share a common ancestor. *J Infect Dis* **196**, 510–521.

Chor, B., Hendy, M. D., Holland, B. R. & Penny, D. (2000). Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Mol Biol Evol* **17**, 1529–1541.

Coulston, J., Naif, H., Brandon, R., Kumar, S., Khan, S., Daniel, R. C. & Lavin, M. F. (1990). Molecular cloning and sequencing of an

Australian isolate of proviral bovine leukaemia virus DNA: comparison with other isolates. *J Gen Virol* **71**, 1737–1746.

Dube, D. K., Sherman, M. P., Saksena, N. K., Bryz-Gornia, V., Mendelson, J., Love, J., Arnold, C. B., Spicer, T., Dube, S. & other authors (1993). Genetic heterogeneity in human T-cell leukemia/lymphoma virus type II. *J Virol* **67**, 1175–1184.

Dube, S., Bachman, S., Spicer, T., Love, J., Choi, D., Esteban, E., Ferrer, J. F. & Poiesz, B. J. (1997). Degenerate and specific PCR assays for the detection of bovine leukaemia virus and primate T cell leukaemia/lymphoma virus *pol* DNA and RNA: phylogenetic comparisons of amplified sequences from cattle and primates from around the world. *J Gen Virol* **78**, 1389–1398.

Dube, S., Dolcini, G., Abbott, L., Mehta, S., Dube, D., Gutierrez, S., Ceriani, C., Esteban, E., Ferrer, J. & Poiesz, B. (2000). The complete genomic sequence of a BLV strain from a Holstein cow from Argentina. *Virology* **277**, 379–386.

Eirin, M. E., Dilernia, D. A., Berini, C. A., Jones, L. R., Pando, M. A. & Biglione, M. M. (2008). Divergent strains of human T-lymphotropic virus type 1 (HTLV-1) within the Cosmopolitan subtype in Argentina. *AIDS Res Hum Retroviruses* **24**, 1237–1244.

Eisen, J. A. (1997). The genetic data environment. A user modifiable and expandable multiple sequence analysis package. *Methods Mol Biol* **70**, 13–38.

Farris, J. S. (1970). Methods for computing Wagner trees. *Syst Zool* **19**, 83–92.

Farris, J. S. (1983). The logical basis of phylogenetic analysis. In *Proceedings of the 2nd meeting of the Willi Hennig Society. Advances in Cladistics 2.* pp. 7–36. Edited by N. F. Platnick & V. A. Funk. New York: Columbia University Press.

Fechner, H., Kurg, A., Geue, L., Blankenstein, P., Mewes, G., Ebner, D. & Beier, D. (1996). Evaluation of polymerase chain reaction (PCR) application in diagnosis of bovine leukaemia virus (BLV) infection in naturally infected cattle. *Zentralbl Veterinarmed B* **43**, 621–630.

Fechner, H., Blankenstein, P., Looman, A. C., Elwert, J., Geue, L., Albrecht, C., Kurg, A., Beier, D., Marquardt, O. & Ebner, D. (1997). Provirus variants of the bovine leukemia virus and their relation to the serological status of naturally infected cattle. *Virology* **237**, 261–269.

Felmer, R., Munoz, G., Zuniga, J. & Recabal, M. (2005). Molecular analysis of a 444 bp fragment of the bovine leukaemia virus gp51 *env* gene reveals a high frequency of non-silent point mutations and suggests the presence of two subgroups of BLV in Chile. *Vet Microbiol* **108**, 39–47.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**, 368–376.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791.

Gatei, M. H., Good, M. F., Daniel, R. C. & Lavin, M. F. (1993). T-cell responses to highly conserved CD4 and CD8 epitopes on the outer membrane protein of bovine leukemia virus: relevance to vaccine development. *J Virol* **67**, 1796–1802.

Gillet, N., Florins, A., Boxus, M., Burteau, C., Nigro, A., Vandermeers, F., Balon, H., Bouzar, A. B., Defoiche, J. & other authors (2007). Mechanisms of leukemogenesis induced by bovine leukemia virus: prospects for novel anti-retroviral therapies in human. *Retrovirology* **4**, 18.

Giribet, G. (2005). TNT: Tree analysis using new technology. *Syst Biol* **54**, 176–178.

Goloboff, P. A. (1993). Estimating character weights during tree search. *Cladistics* **9**, 83–91. doi:10.1080/10635150590905830

Goloboff, P. A. (1999). Analyzing large data sets in reasonable times: solutions for composite optima. *Cladistics* **15**, 415–428. doi:10.1111/j.1096-0031.1999.tb00278.x

Goloboff, P. A. & Farris, J. S. (2001). Methods for quick consensus estimations. *Cladistics* **17**, S26–S34. doi:10.1111/j.1096-0031.2001.tb00102.x

Goloboff, P. A., Farris, J. S. & Nixon, K. C. (2008). TNT, a free program for phylogenetic analysis. *Cladistics* **24**, 774–786. doi:10.1111/j.1096-0031.2008.00217.x

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696–704.

Hasegawa, M., Kishino, H. & Yano, T. (1985). Dating the human–ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**, 160–174.

Hayward, G. S. & Zong, J. C. (2007). Modern evolutionary history of the human KSHV genome. *Curr Top Microbiol Immunol* **312**, 1–42.

Hemmatzadeh, F. (2007). Sequencing and phylogenetic analysis of *gp51* gene of bovine leukaemia virus in Iranian isolates. *Vet Res Commun* **31**, 783–789.

Huelsenbeck, J. P. & Ronquist, F. (2001). MrBayes: Bayesian inference of phylogeny. *Bioinformatics* **17**, 754–755.

Huelsenbeck, J. P., Fredrik Ronquist, F., Nielsen, R. & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310–2314.

Ibrahim, K. M., Nichols, R. A. & Hewitt, G. M. (1996). Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity* **77**, 282–291.

Jones, L. R., Zandomeni, R. & Weber, E. L. (2001). Genetic typing of bovine viral diarrhea virus isolates from Argentina. *Vet Microbiol* **81**, 367–375.

Jones, L. R., Cigliano, M. M., Zandomenni, R. O. & Weber, E. L. (2004). Phylogenetic analysis of bovine pestiviruses: testing the evolution of clinical symptoms. *Cladistics* **20**, 443–453. doi:10.1111/j.1096-0031.2004.00030.x

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**, 3059–3066.

Katoh, K., Kuma, K., Toh, H. & Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* **33**, 511–518.

Kerkhofs, P., Gatot, J. S., Knapen, K., Mammerickx, M., Burny, A., Portetelle, D., Willems, L. & Kettmann, R. (2000). Long-term protection against bovine leukaemia virus replication in cattle and sheep. *J Gen Virol* **81**, 957–963.

Kettmann, R., Couez, D. & Burny, A. (1981). Restriction endonuclease mapping of linear unintegrated proviral DNA of bovine leukemia virus. *J Virol* **38**, 27–33.

Licursi, M., Inoshima, Y., Wu, D., Yokoyama, T., Gonzalez, E. T. & Sentsui, H. (2002). Genetic heterogeneity among bovine leukemia virus genotypes and its relation to humoral responses in hosts. *Virus Res* **86**, 101–110.

Licursi, M., Inoshima, Y., Wu, D., Yokoyama, T., Gonzalez, E. T. & Sentsui, H. (2003). Provirus variants of bovine leukemia virus in naturally infected cattle from Argentina and Japan. *Vet Microbiol* **96**, 17–23.

Makuwa, M., Souquière, S., Clifford, S. L., Telfer, P. T., Sallé, B., Bourry, O., Onanga, R., Mouinga-Ondeme, A., Wickings, E. J. & other authors (2004). Two distinct STLV-1 subtypes infecting Mandrillus sphinx follow the geographic distribution of their hosts. *AIDS Res Hum Retroviruses* **20**, 1137–1143.

Mamoun, R. Z., Morisson, M., Rebeyrotte, N., Busetta, B., Couez, D., Kettmann, R., Hospital, M. & Guillemain, B. (1990). Sequence variability of bovine leukemia virus *env* gene and its relevance to the structure and antigenicity of the glycoproteins. *J Virol* **64**, 4180–4188.

McMahon, M. M. & Sanderson, M. J. (2006). Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. *Syst Biol* **55**, 818–836.

Mirsky, M. L., Olmstead, C. A., Da, Y. & Lewin, H. A. (1996). The prevalence of proviral bovine leukemia virus in peripheral blood mononuclear cells at two subclinical stages of infection. *J Virol* **70**, 2178–2183.

Molteni, E., Agresti, A., Meneveri, R., Marozzi, A., Malcovati, M., Bonizzi, L., Poli, G. & Ginelli, E. (1996). Molecular characterization of a variant of proviral bovine leukaemia virus (BLV). *Zentralbl Veterinarmed B* **43**, 201–211.

Monti, G., Schrijver, R. & Beier, D. (2005). Genetic diversity and spread of *Bovine leukaemia virus* isolates in Argentine dairy cattle. *Arch Virol* **150**, 443–458.

Nixon, K. C. (1999). The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* **15**, 407–414. doi: 10.1111/j.1096-0031.1999.tb00277.x

Nylander, J. A. (2004). MrAIC.pl: Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Phillips, A., Janies, D. & Wheeler, W. (2000). Multiple sequence alignment in phylogenetic analysis. *Mol Phylogenet Evol* **16**, 317–330.

Portetelle, D., Dandoy, C., Burny, A., Zavada, J., Siakkou, H., Gras-Masse, H., Drobecq, H. & Tartar, A. (1989). Synthetic peptides approach to identification of epitopes on bovine leukemia virus envelope glycoprotein gp51. *Virology* **169**, 34–41.

Rambaut, A. & Drummond, A. J. (2007). Tracer v1.4: University of Oxford, Oxford, UK. Available at http://beast.bio.ed.ac.uk/Tracer.

Rannala, B. & Yang, Y. (1996). Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol* **43**, 304–311.

Reichel, M. P., Tham, K. M., Barnes, S. & Kittelberger, R. (1998). Evaluation of alternative methods for the detection of bovine leukaemia virus in cattle. *N Z Vet J* **46**, 140–146.

Rice, N. R., Stephens, R. M., Couez, D., Deschamps, J., Kettmann, R., Burny, A. & Gilden, R. V. (1984). The nucleotide sequence of the env gene and post-env region of bovine leukemia virus. *Virology* **138**, 82–93.

Ronquist, F. & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.

Sagata, N., Yasunaga, T., Tsuzuku-Kawamura, J., Ohishi, K., Ogawa, Y. & Ikawa, Y. (1985). Complete nucleotide sequence of the genome of bovine leukemia virus: its evolutionary relationship to other retroviruses. *Proc Natl Acad Sci U S A* **82**, 677–681.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Simmonds, P. (2006). Recombination and selection in the evolution of picornaviruses and other mammalian positive-stranded RNA viruses. *J Virol* **80**, 11124–11140.

Smith, S. A. & Donoghue, M. J. (2008). Rates of molecular evolution are linked to life history in flowering plants. *Science* **322**, 86–89.

Smith, S. W., Overbeek, R., Woese, C. R., Gilbert, W. & Gillevet, P. M. (1994). The genetic data environment an expandable GUI for multiple sequence analysis. *Comput Appl Biosci* **10**, 671–675.

Swofford, D. L. (1998). PAUP*: Phylogenetic analysis using parsimony (and other methods). Sunderland, MA: Sinauer.

Swofford, D. L., Olsen, G. J., Wadell, P. J. & Hillis, D. M. (1996). Phylogenetic inference. In *Molecular Systematics*, 2nd edn. Edited by D. M. Hillis, C. Moriz & B. K. Mable. Sunderland, MA: Sinauer.

Tartaglia, M., Saulle, E., Pestalozza, S., Morelli, L., Antonucci, G. & Battaglia, P. A. (1998). Detection of bovine mitochondrial DNA in ruminant feeds: a molecular approach to test for the presence of bovine-derived materials. *J Food Prot* **61**, 513–518.

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

Tomaszewski, E. K., Kaleta, E. F. & Phalen, D. N. (2003). Molecular phylogeny of the psittacid herpesviruses causing Pacheco's disease: correlation of genotype with phenotypic expression. *J Virol* **77**, 11260–11267.

Trono, K. G., Perez-Filgueira, D. M., Duffy, S., Borca, M. V. & Carrillo, C. (2001). Seroprevalence of bovine leukemia virus in dairy cattle in Argentina: comparison of sensitivity and specificity of different detection methods. *Vet Microbiol* **83**, 235–248.

Van den Broeke, A., Cleuter, Y., Chen, G., Portetelle, D., Mammerickx, M., Zagury, D., Fouchard, M., Coulombel, L., Kettmann, R. & Burny, A. (1988). Even transcriptionally competent proviruses are silent in bovine leukemia virus-induced sheep tumor cells. *Proc Natl Acad Sci U S A* **85**, 9263–9267.

Vidal, A. U., Gessain, A., Yoshida, M., Tekaia, F., Garin, B., Guillemain, B., Schulz, T., Farid, R. & De Thé, G. (1994). Phylogenetic classification of human T cell leukaemia/lymphoma virus type I genotypes in five major molecular and geographical subtypes. *J Gen Virol* **75**, 3655–3666.

Wiens, J. J. (2006). Missing data and the design of phylogenetic analyses. *J Biomed Inform* **39**, 34–42.

Willems, L., Thienpont, E., Kerkhofs, P., Burny, A., Mammerickx, M. & Kettmann, R. (1993). Bovine leukemia virus, an animal model for the study of intrastrain variability. *J Virol* **67**, 1086–1089.

Willems, L., Burny, A., Collete, D., Dangoisse, O., Dequiedt, F., Gatot, J. S., Kerkhofs, P., Lefèbvre, L., Merezak, C. & other authors (2000). Genetic determinants of bovine leukemia virus pathogenesis. *AIDS Res Hum Retroviruses* **16**, 1787–1795.

Zhao, X. & Buehring, G. C. (2007). Natural genetic variations in bovine leukemia virus envelope gene: possible effects of selection and escape. *Virology* **366**, 150–165.

Zong, J. C., Ciufo, D. M., Alcendor, D. J., Wan, X., Nicholas, J., Browning, P. J., Rady, P. L., Tyring, S. K., Orenstein, J. M. & other authors (1999). High-level variability in the ORF-K1 membrane protein gene at the left end of the Kaposi's sarcoma-associated herpesvirus genome defines four major virus subtypes and multiple variants or clades in different human populations. *J Virol* **73**, 4156–4170.