Classification of small ruminant lentivirus subtype A2, subgroups 1 and 2 based on whole genome comparisons and complex recombination patterns [version 1; peer review: 2 approved with reservations]

Aaron M. Dickey, Timothy P. L. Smith, Michael L. Clawson, Michael P. Heaton, Aspen M. Workman

US Department of Agriculture, Agricultural Research Service, US Meat Animal Research Center, Clay Center, NE, 68933, USA

Abstract

Background: Small ruminant lentiviruses (SRLVs) cause a multisystemic chronic wasting disease in sheep across much of the world. SRLV subtype A2 is prevalent in North America and further classified into multiple subgroups based on variation in the group antigens gene (gag) and envelope (env) genes. In sheep, the ovine transmembrane protein 154 (TMEM154) gene is associated with SRLV susceptibility. Ewes with at least one copy of TMEM154 encoding a full-length protein with glutamate at position 35 (E35; haplotypes 2 and 3), are highly susceptible to SRLV infection while ewes with any combination of TMEM154 haplotypes which encodes lysine (K35; haplotype 1), or truncated proteins (haplotypes 4 and 6) are several times less so. A2 subgroups 1 and 2 are associated with host TMEM154 genotypes; subgroup 1 with the K35/K35 genotype and subgroup 2 with the E35/E35 genotype.

Methods: The goals of this study were to analyze sequence variation within and among SRLV subtype A2 subgroups 1 and 2 and to identify genome-scale recombination patterns. This was done using full-length assemblies of virus samples.

Results: Consensus viral genomes were assembled for 23 infected sheep, including animals of assorted TMEM154 genotypes comprised of haplotypes 1, 2, or 3. Viral genome analysis identified viral subgroups 1 and 2 among the samples, and revealed additional sub-structure within subgroup 2 based on models predicting complex patterns of recombination between the two subgroups in several genomes. Animals with evidence of dual subgroup infection also possessed the most diverse quasi-species and the most highly recombined genomes.

Conclusions: The viral subgroup framework developed to classify
SRLV consensus genomes along a continuum of recombination suggests that animals with the \textit{TMEM154} E35/K35 genotype may represent a reservoir for producing viral genomes representing recombination between A2 subgroups 1 and 2.

**Keywords**
Small ruminant lentivirus, recombination, quasispecies, ovine progressive pneumonia virus
Introduction

Small ruminant lentiviruses (SRLV) are a genetically diverse group of lentiviruses belonging to the family Retroviridae. SRLV infect sheep, goats, and wild ruminants worldwide causing a lifelong persistent infection clinically characterized by wasting, interstitial pneumonia with labored breathing, indurative mastitis, arthritiis, and/or more rarely, encephalitis (Blacklaws, 2012). Disease progression is typically slow, and the genetic background of both the host and virus influence the clinical course and outcome of infection (Heaton et al., 2012; Sider et al., 2013).

The SRLV genome consists of two identical positive-sense single-stranded RNA subunits approximately 9 kb in length. The viral genome, which is integrated into host cells in the form of a provirus, contains three structural genes (gag, pol, and env) and three regulatory genes (tat, vif, and rev) flanked by non-coding long terminal repeat regions (LTRs). Phylogenetic analysis based on partial group antigens gene (gag) and polymerase (pol) gene sequences divides these viruses into five major genotype groups, A-E, which are further divided into different subtypes (Shah et al., 2004). The groups differ by 25–37% and the subtypes differ by 15–27% in nucleotide composition at these genomic loci (Ramírez et al., 2013). Genotypes A and B are distributed worldwide, while genotypes C-E are geographically restricted (Gjerset et al., 2006; Grego et al., 2009; Shah et al., 2004).

The comprehensive set of haplotypes from the same viral species in a single host is known as a quasispecies (Eigen et al., 1988). The viral quasispecies arises from the interplay of three evolutionary factors throughout the duration of a chronic infection. These three factors are mutation, recombination, and selection. Point mutations and small indels are introduced into the SRLV viral genome due to its low fidelity, error prone reverse transcriptase enzyme. Selection can be driven by the host immune system and antiviral medications. Work with related lentiviruses has revealed selection pressure on mutants can produce variants diverging by up to 5% from the founder strain in a few years, though this rate does not remain constant over time (Lee et al., 2008; Shankarappa et al., 1999).

Co-infection (simultaneous infection) or superinfection (sequential infection) can lead to more dramatic genetic changes through recombination, but these two types of dual infection are difficult to distinguish in the absence of a detailed infection history. Regardless of the timing, when diverse viral subtypes infect the same host cell, the reverse transcriptase readily switches between viral genomes, with estimates of three to nine recombination events per replication cycle (Jetzt et al., 2000). This process allows rapid emergence of new viral strains that may exhibit novel phenotypes (reviewed in Ramírez et al., 2013).

In vivo recombination has been documented between genotypes A and B and among genotype A and B subtypes (Andréádóttir, 2003; Fras et al., 2013; Pisoni et al., 2007; Ramírez et al., 2011; Sider et al., 2013). However, recombination in SRLVs has not previously been characterized at the whole-genome level.

The virus-host interaction is a continuous co-evolutionary process. In sheep, genetic variation in the host transmembrane protein 154 (TMEM154) gene associates with SRLV susceptibility (Heaton et al., 2012; Leymaster et al., 2013; Leymaster et al., 2015; Molaee et al., 2018; Molaee et al., 2019; Yaman et al., 2019). Twelve TMEM154 haplotypes have been identified (Heaton et al., 2012), and ewes homozygous for haplotype 1, which encodes a lysine at position 35 (K35), had a decreased risk of SRLV infection compared to those with one or two copies of haplotype 2 or 3 (ancestral), both of which encode a glutamate at position 35 (E35). Two distinct SRLV A2 subgroups have been identified that infected sheep in association with their TMEM154 E35K genotypes and specific diplotypes (Sider et al., 2013). SRLV A2 subgroup 1 viruses were significantly more likely to infect sheep with either TMEM154 diplotype 1,1 or 1,4. TMEM154 haplotype 4 contains a rare frameshift mutation (A4A53) and does not produce mRNA encoding functional amino acids downstream of amino acid position 4 of the precursor protein. Consequently, subgroup 1 associated with hemizygous or homozygous K35 genotypes. SRLV A2 subgroup 2 viruses were more likely to infect sheep with one or two copies of either haplotype 2 or 3, and that could also have one copy of haplotype 4 (Sider et al., 2013). Consequently, subgroup 2 viruses associated with hemizygous or homozygous E35 genotypes. While it has been proposed that TMEM154 E35K hemizygosity or homozgyosity could be a factor in SRLV A2 subgroup associations, the biology remains obscured.

Methods

Ethics statement

All animal procedures were reviewed and approved by the United States Department of Agriculture (USDA), Agricultural Research Service (ARS), U.S. Meat Animal Research Center (USMARC) Animal Care and Use Committee prior to their implementation (Experiment Number 96). The source flock’s history of disease surveillance is also relevant when requesting...
reference samples described in any report. Since first stocking sheep in 1966, USMAC has not had a known case of scrapie. Until 2002, surveillance consisted of monitoring sheep for possible signs of scrapie and submitting brain samples to the USDA Animal and Plant Health Inspection Service (APHIS) National Veterinary Services Laboratory in Ames, IA for testing. All tests have been negative. Since April 2002, USMARC has voluntarily participated in the APHIS Scrapie Flock Certification Program, is in compliance with the National Scrapie Eradication Program, and is certified as scrapie-free. With regards to other transmissible diseases, it is recognized that the USMARC flock of 2000 to 4000 breeding ewes is located in a bluetongue medium incidence area and is known to have some prevalence of contagious ecthyma (sore mouth), foot rot, paratuberculosis (Johne’s disease), ovine progressive pneumonia (visna-maedi), and pseudotuberculosis caseous lymphadenitis.

### Study population

Lungs from 22 sheep at the US Meat Animal Research Center in Nebraska that were a part of the original study for association of A2 subgroups 1 and 2 with *TMEM154* haplotypes (Heaton et al., 2012; Sider et al., 2013) were used in this study (Table 1). These sheep were all genotyped as containing haplotypes 1, 2, or 3. SRLV seropositive sheep were originally diagnosed with clinical ovine progressive pneumonia (OPP) by gross morphology and histopathology of both lung and mediastinal lymph node. In addition, colostrum from one seropositive ewe (201373037) showing no clinical signs of disease was used in this study (Table 1).

### Table 1. Sheep and virus information and for 23 SRLV A2 strains.

<table>
<thead>
<tr>
<th>Animal ID / Viral Strain</th>
<th>GenBank Accession</th>
<th>Breed</th>
<th>TMEM154 DiploTy</th>
<th>Mean Genome Coverage ± Standard Deviation</th>
<th>Submitted Genome Length</th>
<th>Subgroup**</th>
</tr>
</thead>
<tbody>
<tr>
<td>200303038</td>
<td>MT993897</td>
<td>MARCIII</td>
<td>1,2</td>
<td>489.0±426.1</td>
<td>9192</td>
<td>1</td>
</tr>
<tr>
<td>200303332</td>
<td>MT993898</td>
<td>MARCIII</td>
<td>1,1</td>
<td>52.3±20.3*</td>
<td>9171</td>
<td>1</td>
</tr>
<tr>
<td>200303013</td>
<td>MT993896***</td>
<td>MARCIII</td>
<td>1,1</td>
<td>920.0±373.0</td>
<td>9206</td>
<td>1</td>
</tr>
<tr>
<td>200103515</td>
<td>MT993899</td>
<td>MARC III</td>
<td>1,1</td>
<td>71.3±19.4*</td>
<td>9166</td>
<td>1</td>
</tr>
<tr>
<td>200050064</td>
<td>MT993900</td>
<td>ROMANOV-DORSET-SU</td>
<td>1,2</td>
<td>482.8±185.7</td>
<td>9202</td>
<td>r/d</td>
</tr>
<tr>
<td>200323455</td>
<td>MT993901</td>
<td>MARCIII</td>
<td>1,1</td>
<td>1287.9±345.8*</td>
<td>9194</td>
<td>1</td>
</tr>
<tr>
<td>200103342</td>
<td>MT993902</td>
<td>MARCIII</td>
<td>1,1</td>
<td>351.6±124.9</td>
<td>9207</td>
<td>1</td>
</tr>
<tr>
<td>199835918</td>
<td>MT993903</td>
<td>RAMBOUILLET</td>
<td>1,2</td>
<td>1344.8±440.8</td>
<td>9192</td>
<td>1</td>
</tr>
<tr>
<td>200023230</td>
<td>MT993904</td>
<td>MARCIII</td>
<td>1,1</td>
<td>1636.7±1223.3</td>
<td>9193</td>
<td>1</td>
</tr>
<tr>
<td>201373037</td>
<td>MT993905</td>
<td>KATHADIN X RAMBOUILLET</td>
<td>1,1</td>
<td>76.3±31.0*</td>
<td>9166</td>
<td>1</td>
</tr>
<tr>
<td>200117502</td>
<td>MT993906</td>
<td>RAMBOUILLET</td>
<td>1,3</td>
<td>395.8±252.6</td>
<td>9185</td>
<td>r/d</td>
</tr>
<tr>
<td>200216049</td>
<td>MT993907</td>
<td>FINN</td>
<td>1,3</td>
<td>540.2±280.9</td>
<td>9200</td>
<td>r/d</td>
</tr>
<tr>
<td>200212120</td>
<td>MT993908</td>
<td>POLLED DORSET</td>
<td>1,2</td>
<td>927.8±611.7</td>
<td>9202</td>
<td>r/d</td>
</tr>
<tr>
<td>200312013</td>
<td>MT993909</td>
<td>POLLED DORSET</td>
<td>2,2</td>
<td>651.0±410.0</td>
<td>9191</td>
<td>r/d</td>
</tr>
<tr>
<td>200312088</td>
<td>MT993910</td>
<td>POLLED DORSET</td>
<td>1,2</td>
<td>279.3±195.5</td>
<td>9199</td>
<td>r/d</td>
</tr>
<tr>
<td>199906011</td>
<td>MT993911***</td>
<td>TEXEL</td>
<td>2,2</td>
<td>692.7±417.4</td>
<td>9189</td>
<td>2a</td>
</tr>
<tr>
<td>200106932</td>
<td>MT993912</td>
<td>TEXEL</td>
<td>2,3</td>
<td>1230.6±1086.1</td>
<td>9191</td>
<td>2a</td>
</tr>
<tr>
<td>200106929</td>
<td>MT993913</td>
<td>TEXEL</td>
<td>2,2</td>
<td>580.3±521.8</td>
<td>9201</td>
<td>2a</td>
</tr>
<tr>
<td>200016283</td>
<td>MT993914</td>
<td>FINN</td>
<td>1,2</td>
<td>2014.5±710.3</td>
<td>9206</td>
<td>2b</td>
</tr>
<tr>
<td>200335185</td>
<td>MT993915</td>
<td>RAMBOUILLET</td>
<td>1,2</td>
<td>171.8±74.7</td>
<td>9203</td>
<td>2b</td>
</tr>
<tr>
<td>200177363</td>
<td>MT993916</td>
<td>DORSET X ROMANOV</td>
<td>1,3</td>
<td>1801.5±1564.7</td>
<td>9215</td>
<td>2b</td>
</tr>
<tr>
<td>199916128</td>
<td>MT993917</td>
<td>FINN</td>
<td>1,1</td>
<td>227.6±67.3*</td>
<td>9172</td>
<td>2b</td>
</tr>
<tr>
<td>199916193</td>
<td>MT993918</td>
<td>FINN</td>
<td>1,1</td>
<td>2661.1±1329.2</td>
<td>9204</td>
<td>2b</td>
</tr>
</tbody>
</table>

*No long reads: coverage includes only short reads, 10 to 26 nucleotides not fully resolved on 5’ ends.

**The population to which a consensus genome was assigned in the subgroup 1 vs 2a vs 2b recombination model (Figure 3A).

r/d: recombinant/dual infections.

***The short read + long read consensuses reported here are identical (200303013) and slightly different (199906011) from the long-read consensuses reported by Workman et al., 2017 (KY358787 and KY358788 respectively). See results.
Generation of full-length SRLV genomes

Lung samples were homogenized using a gentleMACS dissociator (Miltenyl Biotec; San Diego, CA) in minimal essential medium (Gibco, Thermo Fisher Scientific, Waltham, MA). Homogenates were then subjected to two freeze/thaw cycles to further ensure cell lysis. Homogenates were clarified by centrifugation followed by sequential filtration through 0.45 and 0.2-μm syringe filters to remove cellular debris. Clarified samples (250 μL) were treated with 20 U RNase One (Promega, Madison, WI) and 30 U Turbo DNase (Ambion, Austin, TX) in 1x DNase buffer (Ambion) at 37°C for 90 minutes to degrade unprotected host and environmental nucleic acids. To ensure continuous DNase activity, 10 U of DNase was added to the sample every 30 minutes during the 90-minute incubation. Remaining nucleic acids were isolated using Trizol LS (Invitrogen, Carlsbad, CA) according to the manufacturer’s instructions. A final DNase treatment was performed to remove final traces of DNA from the RNA preparation.

Colostrum (approximately 4 mL) was manually collected from a SRLV seropositive ewe within the first 24 hours after giving birth. Colostrum was diluted 1:2 with cold phosphate-buffered saline and centrifuged at 800 x g for 15 minutes at 4°C. The cream layer was skimmed from the top and 250 μL of milk was treated with nucleases and RNA was extracted as described above.

RNA libraries were prepared as previously described (Workman et al., 2015; Workman et al., 2017; Workman et al., 2018). Briefly, 100 ng of total RNA were used as input material for Illumina TruSeq RNA Sample Preparation Kit (Illumina, San Diego, CA). Libraries were prepared as specified by the manufacturer’s protocol without the initial step of poly-A selection on oligo-dT beads. Libraries were sequenced on an Illumina MiSeq instrument with a 600-cycle kit (v3) to generate 2 x 300 bp paired-end reads. Index adapters were removed from raw sequence reads using cutadapt 1.9.1 (Martin, 2011) or BBduk 35.82 (Brian Bushnell within Geneious 11.1.4) (Kearse et al., 2012) (Biomatters, Auckland, New Zealand) and trimmed reads were screened against the UniVec_Core database to remove contaminating vector sequences. Overlapping paired-end reads were merged using BBMerge 8.9 (Bushnell within Geneious).

The remaining RNA was used to generate long-read sequencing libraries according to a modified RNA Iso-Seq with poly(A) tails added to the 3’ ends to allow cDNA synthesis of subgenomic fragments. The resulting cDNA was amplified, size fractionated and the largest fragments were used to make SMRTbell templates, which were sequenced on a Pacific Biosciences RSII instrument. SMRT Analysis was used to generate error corrected circular consensus sequences (CCS) from the raw reads and adapters and poly(A) tails were removed with BBduk.

Reads greater than 1,000 nucleotides in length were assembled de novo with the Geneious assembler. All reads were then mapped to the de novo assembly, and the resulting consensus sequence was reported. Four strains lacked good quality long-read sequence data (Table 1) so the MiSeq short reads were assembled using template-assisted assembly in Geneious following Workman et al. (2018) with accessions KY358787 and KY358788, respectively, used as subgroup 1 and 2 references. These two reference strains were included in this study.

To calculate total genome coverage for each sample, merged and unmerged paired-end reads plus long-read CCS’s were jointly mapped to the consensus genome using the Geneious mapper with 40% maximum allowable mismatches, a word length of 24, an index word length of 14, 10% allowable gaps and a maximum gap size of 50.

Up to 26 nucleotides on the 5’ ends did not fully resolve in genomes with only short-read sequencing available (Table 1). Genomes were manually annotated based on alignments with full-length SRLV genomes available in GenBank.

Phylogenetic analyses of full-length genomes

To augment the newly reported genome sequences, full-length SLRV and SLRV-like genomes were also downloaded from GenBank using the following query of the Nucleotide database on October 11, 2019: txid11660[ORGN] OR txid2169971[ORGN] OR txid11653[ORGN] OR txid11663[ORGN] AND (~“8000”[SLEN] : “12000”[SLEN]) OR txid11660[ORGN] OR txid11653[ORGN] OR txid11663[ORGN] AND (~“8000”[SLEN] : “12000”[SLEN]). 79 unique genomes were aligned using MUSCLE 3.8.425 (Edgar, 2004 in Geneious 11.1.5) and a neighbor net phylogenetic network was built using default settings in Splitstree5 (Huson & Bryant, 2006).

A2 Subgroup diagnostic SNP identification accounting for recombination

Population assignment of each genome generated in this study was performed in FineStructure version 4 and its companion program, ChromoPainter version 1 (Lawson et al., 2012). To generate the alignment used in FineStructure, MUSCLE was used followed by manual refinement. Gaps in the alignment were converted to presence/absence characters as-is with simple gap patterns reduced to a single character regardless of size. The recombination rate map used in FineStructure and Chromopainter was estimated using the LDhat 2.2a interval program (McVean et al., 2004). For LDhat, pre-computed likelihoods were generated using a population-scaled per-site mutation rate inferred from the data (0.07587), a grid size of 101 and a maximum population-scaled whole-genome recombination rate of 100. The variable rate estimation was run for 10 million iterations with the first half discarded as burn-in and a block penalty of 20. To avoid alignment edge inaccuracies, the final 10% of SNPs from the 5’ terminus were placed preceding the 5' terminus and the first 10% of SNPs from the 5' terminus were placed following the 3' terminus, essentially simulating circular genomes. The recombination rate point estimates at these simulated edges were removed. The outputs of LDhat were population-scaled recombination rates (r), which relate to the biochemical recombination rate (r) according to the formula \( p = 2N_e r \) where \( N_e \) is the effective population size. \( N_e \) is difficult to estimate. Estimates for HIV, a related lentivirus which also produces chronic infections, vary by several orders
of magnitude (Pennings et al., 2014). Computational estimates of $N_e$ in viruses also require time-series sampling data (Rousseau et al., 2017). Thus, $N_e$ was not estimated for this study and the ChromoPainter recombination outputs were interpreted as being scaled by $2N_e$. The FineStructure analysis was run using the linkage model, the variable recombination rate map estimated as described above, and specifications for haploid genomes. ChromoPainter detected shared ancestry by reconstructing each genome as a probabilistic mosaic of ‘chunks’ derived via recombination from all other input genomes (termed ‘all vs all painting’) and FineStructure assigned the genomes to populations based on the quantity and lengths of these shared genomic chunks. The following settings were changed from the default in FineStructure and/or ChromoPainter to ensure convergence of estimated parameters: ChromoPainter chunks-per-region parameter $k=38$, ChromoPainter samples $s=10$, Markov Chain Monte Carlo (MCMC) iterations $i=1e6$ FineStructure tree finding maximization steps $j=1e6$ and FineStructure independent MCMC runs $r=10$. ChromoPainter iterations $i=100$ were run for estimating the global switch rate parameter and the global mutation rate. In FineStructure, Strain 201373037 was excluded from the estimation of the global switch rate parameter since it trended toward 0 and stalled the program. This indicated very closely related samples in the dataset (G. Hellenthal, personal communication). Inconsistency in assignment of individuals to populations was resolved by assigning all ambiguously assigned individuals to the largest of the potential populations.

To model subgroup-specific recombination, ChromoPainter version 2 was run in ‘donor-mode’ using the population assignments, global switch-rate parameter $(0.168355)$, and population specific mutation rates output from the FineStructure analysis. These models were run for 500 iterations to ensure convergence of copy probability. In contrast to all vs all painting, donor specific painting assigns genomic chunks to recipient genomes based on donor populations comprised of multiple genomes. To increase genome-wide assignment probability of subgroup specific SNPs, consensus genomes with evidence of large recombination blocks were iteratively removed from each subgroup pool of donor genomes if average copy probability was increasing. This was done to eliminate the most obvious recombinants from the pools while retaining enough donor genomes to optimize the recombination model. The output donor subgroup-assignments for each SNP were used to identify sub-group-specific SNPs while accounting for recombination. Recombination rates have been estimated for several RNA retroviruses, and most estimates are in the range of $10^{-3}$ to $10^{-5}$ (Thomas et al., 2014). Thus, we also ran the diagnostic SNP identification model using a range of fixed recombination rates ($10^{-3}$ to $10^{-8}$ Morgans-per-base pair) to confirm that diagnostic SNP count did not change when varying input recombination rate by several orders of magnitude.

**Dual infection inference using diagnostic SNPs**

The subgroup-specific SNP content was quantified for each strain by extracting intra-host SNPs meeting default statistical restrictions (Maximum Variant P-value of $10^{-5}$; Minimum Strand-Bias P-value of $10^{-4}$ when exceeding 65% bias) relative to their final alignment in Geneious 11.1.4. The percentage of subgroup 1, subgroup 2, and ‘other’ SNPs at each diagnostic locus was calculated for each consensus genome. Subgroup partial dual infections were inferred as contiguous or nearly contiguous SNP blocks bearing both subgroup diagnostic alleles. For visualization relative to the consensus genome, these inferred partial dual infections were limited to genome blocks or scaffolded $\geq50$ nucleotides long where variants diagnostic for both subgroups co-occurred at a frequency of $\geq5\%$ and $\geq2$ reads. The 5% threshold was chosen as it was a conservative value accounting for sequencing error and mis-mapping when calling quasispecies SNPs. Multiple putative dual infection blocks were extended or scaffolded together when separated by $<50$ nucleotides and 1 diagnostic SNP. We characterize these as being caused by dual infection with unknown underlying viral haplotypes containing SNPs diagnostic to both subgroups at these regions as this is the most parsimonious explanation. However, quasispeciation in the absence of dual infection is an increasingly possible explanation as the numbers of adjacent subgroup diagnostic SNPs in the characterized regions decrease.

**Functional analyses and annotation**

Once subgroup-specific SNPs were identified in the context of recombination, intra-host amino acid variation (functional quasispecies) at the subgroup diagnostic loci were identified by extracting variants from the alignments in Geneious 11.1.5 using the same statistical criteria applied to nucleotides and occurring at a frequency of $\geq5\%$. Highly variable domains in gag and env previously shown to be important were analyzed in the context of subgroup assignment and host TMEM154 diplotype. Additionally, SignalP-5.0 (Nielsen, 2017) was used to predict the env signal peptide cleavage site.

**Results**

**Genomes**

Coverage of the 23 genomes ranged from 52- to 2661-fold (Table 1). Complete or near-complete genomes ranged from 9164 to 9215 nucleotides in length. The combined short read and long read consensus genome of strain 199906011 was slightly different from the long read only consensus (KY358788). The sites with differences had a high frequency of the minor allele in the quasispecies in the long read only consensus and switched the identity of the minor allele in the combined short read and long read consensus. The combined short read and long read consensus of strain 200303013 was identical to the long read only consensus (KY358787). A phylogenetic network using full-length SRLV genomes was dominated by groups A and B (Figure 1). Subtype A2 strains from the United States of America occupied a distinct cluster within the network. Additional clusters on the tree were generally represented by a single geographic region, with Italy representing the highest number of unique clusters. Recombination was evident in several clusters of the network including subtype A2. FineStructure analysis of the 23 genomes from this study identified six distinct populations across the consensus genomes based on all-vs-all genome painting (Figure 2). FineStructure identified two distinct subgroup 2 populations identified in Figure 2 as subgroup 2a and subgroup 2b. Subgroup 2a is intermediate...
Figure 1. Splitstree neighbornet phylogenetic network of 79 SRLV genomes from genotypes A, B, C and E. Colors correspond to genomes assigned to subgroup specific pools of donor genomes in recombination analysis (Table 1).
between subgroup 1 and subgroup 2b on principal component 1 (Figure 2).

Recombination models
Due to the large number of FineStructure identified populations, several recombination models were run in ChromopainterV2. Five of the six identified populations had >1 consensus genome. The model with five potential populations of donor genomes indicated that the three most frequent populations (subgroup 1, subgroup 2a and subgroup 2b) accounted for >88% of the SNPs and were the majority donors to 22 genomes (Extended data, Supplementary Figure 1 (Dickey & Workman, 2020b)). The model was run with these three donor populations only (Figure 3A). Due to the relative location of the four populations along principal component 1 (Figure 2), the model was also run with subgroup 2b and subgroup 2a as the only two donor populations (Figure 3B). This showed possible complex recombination blocks between subgroup 1 and subgroup 2b in subgroup 2a genomes (Figure 3B). Subgroup 2a genomes also showed intermediate average pairwise percent divergences between subgroups 1 and 2b (Table 3). All models showed many predicted recombination blocks spanning the consensus genomes (Table 2, Figure 3 and Extended data, Supplementary Figure 1 (Dickey & Workman, 2020b)). Finally, a recombination model was run with only 2 donor populations, subgroup 1 and subgroup 2 (as 2a+2b) (Figure 4). This was done to identify and extract subgroup specific SNPs while accounting for recombination. Viral strains 200050064 and 199916128 were removed from the subgroup 1 and 2 donor pools respectively based on having the highest proportion of inter-subgroup recombination (Table 2, Figure 3) and this improved the subgroup 1 vs subgroup 2 recombination model. Further removal of genomes as potential donors did not improve the model. The average number of alternate subgroup recombination blocks (Chromopainter’s chunkcount) was 2-fold higher in subgroup 2 genomes than subgroup 1 (Table 2). Chromopainter’s chunklengths parameter averaged
Figure 3. Twenty-three SRLV subtype A2 genomes ‘painted’ with recipient genomic ‘chunks’ derived from subgroup-specific donor genomes. The recombination models utilized were (A) subgroup 1 vs subgroup 2a vs subgroup 2b and (B) subgroup 1 vs subgroup 2b. The black boxes highlight large subgroup 1 recombination blocks in subgroup 2a genomes.
3-fold higher in subgroup 2 and predicted population specific mutation rate averaged 3-fold higher in subgroup 2 consensus genomes than subgroup 1 (Table 2).

Subgroup diagnostic SNP inference accounting for recombination between subgroup 1 and subgroup 2 (as FineStructure population 2a+2b) resulted in 413 diagnostic SNPs (Extended Table 2).

**Table 2.** Select ChromopainterV2 calculated parameters for the subgroup 1 vs 2 recombination model (Figure 4). Chromopainter's 'chunkcounts' parameter is defined as the number of genomic chunks from a population of donor genomes, assigned to the recipient genome via recombination. 'Chunklengths' are the combined lengths (in centimorgans X 2Ne) of these chunks. Donor specific mutation rate is the amount of mismatching across the recipient chunklengths divided by the number of loci. Donor status is the population of donor genomes to which a consensus genome was assigned in the final subgroup 1 vs subgroup 2 recombination model (Figure 4). To improve recombination models, genomes with high inter-subgroup recombination were iteratively removed from the populations if model quality (as judged by average copy probability) was increasing.

<table>
<thead>
<tr>
<th>Recipient Genome</th>
<th>FineStructure Assigned Population</th>
<th>Donor Status</th>
<th>Subgroup 1 Chunk counts</th>
<th>Subgroup 2 Chunk counts</th>
<th>Subgroup 1 Chunk lengths</th>
<th>Subgroup 2 Chunk lengths</th>
<th>Donor specific mutation rate*</th>
</tr>
</thead>
<tbody>
<tr>
<td>200303038</td>
<td>Subgroup 1</td>
<td>Subgroup 1</td>
<td>52.4404</td>
<td>0</td>
<td>30218.2</td>
<td>0</td>
<td>0.0161771</td>
</tr>
<tr>
<td>200303332</td>
<td>Subgroup 1</td>
<td>Subgroup 1</td>
<td>51.8010</td>
<td>0</td>
<td>30218.2</td>
<td>0</td>
<td>0.0177648</td>
</tr>
<tr>
<td>200303013</td>
<td>Subgroup 1</td>
<td>Subgroup 1</td>
<td>60.0599</td>
<td>0</td>
<td>30218.2</td>
<td>0</td>
<td>0.0357224</td>
</tr>
<tr>
<td>200103515</td>
<td>Subgroup 1</td>
<td>Subgroup 1</td>
<td>60.6890</td>
<td>0</td>
<td>30218.2</td>
<td>0</td>
<td>0.0236503</td>
</tr>
<tr>
<td>200050064</td>
<td>Subgroup 1</td>
<td>Recipient only</td>
<td>48.2632</td>
<td>6.4329</td>
<td>28003.9</td>
<td>2214.3</td>
<td></td>
</tr>
<tr>
<td>200323455</td>
<td>Subgroup 1</td>
<td>Subgroup 1</td>
<td>55.0174</td>
<td>0</td>
<td>30218.2</td>
<td>0</td>
<td>0.0196987</td>
</tr>
<tr>
<td>200103342</td>
<td>Subgroup 1</td>
<td>Subgroup 1</td>
<td>60.1737</td>
<td>0.0740</td>
<td>30207.5</td>
<td>10.7</td>
<td>0.0160127</td>
</tr>
<tr>
<td>199835918</td>
<td>Subgroup 1</td>
<td>Subgroup 1</td>
<td>61.1799</td>
<td>0</td>
<td>30218.2</td>
<td>0</td>
<td>0.0294231</td>
</tr>
<tr>
<td>200023230</td>
<td>Subgroup 1</td>
<td>Subgroup 1</td>
<td>62.7606</td>
<td>0</td>
<td>30218.2</td>
<td>0</td>
<td>0.0128684</td>
</tr>
<tr>
<td>201373037</td>
<td>Subgroup 1</td>
<td>Subgroup 1</td>
<td>46.9864</td>
<td>0</td>
<td>30218.2</td>
<td>0</td>
<td>0.0480409</td>
</tr>
<tr>
<td>200117502</td>
<td>Pop 5</td>
<td>Recipient only</td>
<td>25.5460</td>
<td>16.0997</td>
<td>15790.2</td>
<td>14428.0</td>
<td></td>
</tr>
<tr>
<td>200216049</td>
<td>Pop 5</td>
<td>Recipient only</td>
<td>42.0507</td>
<td>43.0426</td>
<td>15384.2</td>
<td>14834.0</td>
<td></td>
</tr>
<tr>
<td>200212120</td>
<td>Pop 4</td>
<td>Recipient only</td>
<td>43.3061</td>
<td>36.3817</td>
<td>18021.3</td>
<td>12196.9</td>
<td></td>
</tr>
<tr>
<td>200312013</td>
<td>Pop 4</td>
<td>Recipient only</td>
<td>39.8241</td>
<td>22.9727</td>
<td>21460.6</td>
<td>8757.6</td>
<td></td>
</tr>
<tr>
<td>200312088</td>
<td>Pop 6</td>
<td>Recipient only</td>
<td>27.2590</td>
<td>65.5737</td>
<td>8656.4</td>
<td>21561.9</td>
<td></td>
</tr>
<tr>
<td>199906011</td>
<td>Subgroup 2a</td>
<td>Subgroup 2</td>
<td>2.8228</td>
<td>47.0027</td>
<td>582.9</td>
<td>29635.3</td>
<td>0.0628346</td>
</tr>
<tr>
<td>200106932</td>
<td>Subgroup 2a</td>
<td>Subgroup 2</td>
<td>1.0739</td>
<td>38.8866</td>
<td>469.4</td>
<td>29748.8</td>
<td>0.0525776</td>
</tr>
<tr>
<td>200106929</td>
<td>Subgroup 2a</td>
<td>Subgroup 2</td>
<td>1.4457</td>
<td>44.8749</td>
<td>339.5</td>
<td>29878.8</td>
<td>0.0684202</td>
</tr>
<tr>
<td>200016283</td>
<td>Subgroup 2b</td>
<td>Subgroup 2</td>
<td>0</td>
<td>63.3829</td>
<td>0</td>
<td>30218.2</td>
<td>0.0781928</td>
</tr>
<tr>
<td>200335185</td>
<td>Subgroup 2b</td>
<td>Subgroup 2</td>
<td>0</td>
<td>66.1020</td>
<td>0</td>
<td>30218.2</td>
<td>0.0929781</td>
</tr>
<tr>
<td>200177363</td>
<td>Subgroup 2b</td>
<td>Subgroup 2</td>
<td>0</td>
<td>70.5193</td>
<td>0</td>
<td>30218.2</td>
<td>0.0898418</td>
</tr>
<tr>
<td>199916128</td>
<td>Subgroup 2b</td>
<td>Subgroup 2</td>
<td>0</td>
<td>70.2713</td>
<td>0</td>
<td>30218.2</td>
<td>0.0938431</td>
</tr>
<tr>
<td>199916193</td>
<td>Subgroup 2b</td>
<td>Recipient only</td>
<td>6.1441</td>
<td>53.8583</td>
<td>5726.2</td>
<td>24492.0</td>
<td>-</td>
</tr>
</tbody>
</table>

*ChromopainterV2 mutationprobs parameter.
data, Supplementary Table 1 (Dickey & Workman, 2020a)). The frequency of alternate subgroup diagnostic alleles was 3-fold higher in subgroup 2 consensus genomes than subgroup 1 (Table 4).

Intra-host variation at subgroup-specific SNPs was analyzed to infer the presence of dual infections. The parameters specified for predicting subgroup dual infections resulted in 73 genomic regions indicative of dual infection across nine consensus genomes (range: 1–14, average: 8.2), averaging 261.6 nucleotides in length (range: 55–1482) and comprising 2%–45% of the genome (Extended data, Supplementary Table 2 (Dickey & Workman, 2020a), Figure 5).

Functional variation

Of the 413 subgroup diagnostic SNPs identified, 106 were nonsynonymous (Extended data, Supplementary Tables 1 and 3 (Dickey & Workman, 2020a)). A2 subgroup 1 and 2 specific variants were identified in all viral genes with frequencies ranging from 2.2 to 4.3%. Sequence analysis of the immunodominant epitope in the gag gene revealed two adjacent SNPs that resulted in a single amino acid change distinguishing subgroups 1 and 2 (Extended data, Supplementary Table 3 (Dickey & Workman, 2020a)). Analysis of the env gene variable regions V1-V5 (Valas et al., 2000) found no subgroup specific SNP in variable regions V1 and V2, five subgroup specific variants each in V3 and V4, and one in V5 (Extended data, Supplementary

---

### Table 3. Average pair-wise genetic distance within (on-diagonal) and among (off-diagonal) SLRV A2 subgroups characterized in this study.

<table>
<thead>
<tr>
<th></th>
<th>Subgroup 1</th>
<th>Subgroup 2a</th>
<th>Subgroup 2b</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgroup 1</td>
<td>1.7%</td>
<td>13.7%</td>
<td>15.6%</td>
</tr>
<tr>
<td>Subgroup 2a</td>
<td>13.7%</td>
<td>5.1%</td>
<td>10.4%</td>
</tr>
<tr>
<td>Subgroup 2b</td>
<td>15.6%</td>
<td>10.4%</td>
<td>5.2%</td>
</tr>
</tbody>
</table>

---

**Figure 4.** Twenty-three SRLV subtype A2 genomes ‘painted’ with recipient genomic ‘chunks’ derived from subgroup-specific donor genomes. The recombination model utilized was subgroup 1 vs subgroup 2.
Table 4. Subgroup specific intra-host genetic variation (quasispecies) for 23 SRLV A2 consensus genomes. Donor status is the population of donor genomes to which a consensus genome was assigned in the final subgroup 1 vs subgroup 2 recombination model (Figure 4).

<table>
<thead>
<tr>
<th>Strain</th>
<th>Donor Status</th>
<th>Average % Subgroup 1 Across 413 Diagnostic Loci</th>
<th>Average % Subgroup 2 Across 413 Diagnostic Loci</th>
<th>Average % ‘Other’ Across 413 Diagnostic Loci</th>
<th>Genomic Regions Indicative of Dual Infection</th>
<th>Total Length of Genomic Regions Indicative of Dual Infection</th>
</tr>
</thead>
<tbody>
<tr>
<td>200303038</td>
<td>Subgroup 1</td>
<td>99.79%</td>
<td>0.16%</td>
<td>0.05%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200303332</td>
<td>Subgroup 1</td>
<td>99.89%</td>
<td>0.02%</td>
<td>0.08%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200303013</td>
<td>Subgroup 1</td>
<td>99.55%</td>
<td>0.17%</td>
<td>0.28%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200103515</td>
<td>Subgroup 1</td>
<td>97.18%</td>
<td>1.75%</td>
<td>1.08%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200050064</td>
<td>Recipient only</td>
<td>91.70%</td>
<td>7.74%</td>
<td>0.56%</td>
<td>6</td>
<td>1427</td>
</tr>
<tr>
<td>200323455</td>
<td>Subgroup 1</td>
<td>99.25%</td>
<td>0.43%</td>
<td>0.32%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200103342</td>
<td>Subgroup 1</td>
<td>98.93%</td>
<td>0.73%</td>
<td>0.34%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>199835918</td>
<td>Subgroup 1</td>
<td>98.89%</td>
<td>0.62%</td>
<td>0.49%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200023230</td>
<td>Subgroup 1</td>
<td>98.99%</td>
<td>0.50%</td>
<td>0.51%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>201373037</td>
<td>Subgroup 1</td>
<td>98.78%</td>
<td>0.71%</td>
<td>0.51%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200117502</td>
<td>Recipient only</td>
<td>33.22%</td>
<td>66.24%</td>
<td>0.70%</td>
<td>9</td>
<td>3394</td>
</tr>
<tr>
<td>200216049</td>
<td>Recipient only</td>
<td>52.04%</td>
<td>47.62%</td>
<td>0.34%</td>
<td>13</td>
<td>3704</td>
</tr>
<tr>
<td>200212120</td>
<td>Recipient only</td>
<td>72.61%</td>
<td>25.48%</td>
<td>1.91%</td>
<td>11</td>
<td>4179</td>
</tr>
<tr>
<td>200312013</td>
<td>Recipient only</td>
<td>77.65%</td>
<td>21.69%</td>
<td>0.66%</td>
<td>8</td>
<td>1798</td>
</tr>
<tr>
<td>200312088</td>
<td>Recipient only</td>
<td>40.74%</td>
<td>58.20%</td>
<td>1.07%</td>
<td>14</td>
<td>2297</td>
</tr>
<tr>
<td>199906011</td>
<td>Subgroup 2</td>
<td>4.25%</td>
<td>95.12%</td>
<td>0.63%</td>
<td>1</td>
<td>158</td>
</tr>
<tr>
<td>200106932</td>
<td>Subgroup 2</td>
<td>1.00%</td>
<td>98.18%</td>
<td>0.82%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200106929</td>
<td>Subgroup 2</td>
<td>0.92%</td>
<td>98.49%</td>
<td>0.58%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200016283</td>
<td>Subgroup 2</td>
<td>0.93%</td>
<td>98.67%</td>
<td>0.40%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200335185</td>
<td>Subgroup 2</td>
<td>0.35%</td>
<td>99.27%</td>
<td>0.37%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>200177363</td>
<td>Subgroup 2</td>
<td>1.62%</td>
<td>97.60%</td>
<td>0.79%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>199916128</td>
<td>Subgroup 2</td>
<td>1.92%</td>
<td>96.40%</td>
<td>1.68%</td>
<td>2</td>
<td>241</td>
</tr>
<tr>
<td>199916193</td>
<td>Recipient only</td>
<td>19.47%</td>
<td>79.90%</td>
<td>0.62%</td>
<td>9</td>
<td>1896</td>
</tr>
</tbody>
</table>

Table 3 (Dickey & Workman, 2020a)). Six subgroup defining variants were identified in the predicted env signal peptide.

Discussion

This study provides full-length or near-full-length consensus genomes from 21 new SRLV subtype A2 strains used in determining the viral subgroup association with TMEM154 E35K genotypes (Heaton et al., 2012; Sider et al., 2013) in addition to the two subgroup representative strains from Workman et al., 2017. These genomes were analyzed for recombination and population structure using a chromosome ‘painting’ model. Several genomes showed complex recombination patterns. Furthermore, this model was used to identify 413 subgroup-specific SNPs while accounting for recombination. This information was used to quantify intra-host genetic diversity at diagnostic SNPs and estimated nine animals were dually infected with viral recombinants such that they have diagnostic SNPs from both virus subgroups for portions of their genome. Lastly, we analyzed important functional domains in the virus genome in the context of virus subgroup and host TMEM154 diplotypes focusing only on haplotypes 1, 2 and 3.

The SRLV phylogenetic network contained subtype A2 as a distinct cluster (Figure 1). Several genomes in this cluster are connected by many nodes indicating inter-subgroup recombination. The FineStructure analysis identified two distinct subgroup 2 ‘populations’ of consensus genomes. These have been provisionally designated subgroup 2a and subgroup 2b.
Subgroup 2a is intermediate between subgroup 1 and subgroup 2b both in terms of position along principal component 1 (Figure 2) and in terms of genetic distance (Table 3). The subgroup 2 genome reported by Workman et al., 2017 (viral strain 199906011, GenBank KY358788.1) belongs to subgroup 2a (Table 1). This is a case where it is difficult to distinguish, with certainty, the recombinant from the second viral donor, however there is some evidence that subgroup 2b may be a ‘purer’ representative of subgroup 2. Even when subgroup 2a genomes were included among the subgroup 2 donors, the only consensus genomes that resolved unambiguously as subgroup 2 were the four subgroup 2b strains (Table 2). In contrast, subgroup 1 was more clearly delimited by the recombination model. Of the nine subgroup 1 donor genomes, only 200103342 did not resolve unambiguously as subgroup 1 (Table 2).

Subgroup 2 was more genetically diverse than subgroup 1 based on its higher mutation rates and increased recombination (Figure 4, Table 2). Subgroup 2 also had higher intra-host genetic diversity based on the dual infection analysis (Table 4). The consensus genomes with the highest intra-host diversity (Figure 5, Table 4 and Extended data, Supplementary Table 2 (Dickey & Workman, 2020a)) and highest recombination block count (Figure 4, Table 2) did not conform to a good concept of ‘population’ (Extended data, Supplementary Figure 1 (Dickey & Workman, 2020b)) despite being identified as populations by FineStructure analysis (Figure 2). But these genomes could be parsimoniously modeled as complex recombinants of subgroups 1 and 2 (Figure 4). We argue that more than one representative genome is required to properly distinguish subgroups from recombinant forms of subgroups 1 and 2. The subgroup 1 vs subgroup 2 recombination model used to identify diagnostic SNPs was informed by both recombination rate variation across the genome as well as differing mutation rates between subgroups. Models run with a constant recombination rate across the genome identified a slightly higher number (421) of diagnostic SNPs (see Methods). The dual infection analysis identified multiple genomic regions featuring modest frequencies of both subgroup diagnostic alleles (Extended data, Supplementary Table 2 (Dickey & Workman, 2020a)). These were identified in the most recombinant consensus genomes (Figure 4). Because the dual infection regions did not span the entire genome, the underlying haplotypes were most likely recombinant as opposed to ‘pure’ subgroup sequences.

The relatively small number of strains characterized in this study and a paucity of geographic variability biased our results and limited our ability to make recombination-based inference.
While the subset of samples from the original TMEM154-A2 association studies chosen was a good starting place for modeling recombination, the addition of larger numbers of geographically diverse SRLV A2 genomes should improve the recombination model(s) substantially due to a larger pool of potential recombinant and parental genomes (Yahara et al., 2019). There are presently 79 unique full length SRLV genomes available with more than half of these published since 2019 (Colitti et al., 2019; present work) so the time has probably come to recharacterize SRLV diversity at the whole-genome level, expanding the current classification beyond partial gag/pol sequence. A revised classification system will better facilitate outbreak tracing and identification of recombinants circulating beyond their local flocks. Such circulating recombinant forms (CRFs) have been extensively characterized for HIV (Carr et al., 1999; Leitner et al., 2005) providing a possible model and framework for the SRLV research community to adapt. However, the current CRF framework for HIV utilizes consensus genomes so an accounting of the underlying haplotypes (quasispecies) contributing to these consensuses would benefit the genomic characterization of both lentiviruses.

Fras et al., 2013 have suggested that dual infection of small ruminant lentiviruses may be common, understudied, and under-reported. Our results and those of Sider et al., 2013 confirm that dual infection is common though none of our samples showed evidence of having been dually infected by pure subgroup 1 and 2 representatives. Hopefully, declining sequencing costs and increased availability of whole genome sequencing will foster greater reporting of this phenomenon. Our results also conform to those of Sider et al., 2013 including the two subgroups identified and the existence of recombination. These results extend those of Sider et al., 2013 from partial gag/env to the complete genome while accounting for recombination. Recombination is also clearly delimited by the ChromoPainter models. The largest recombination blocks were also predicted by the software program, RDP (Martin et al., 2015) (data not shown), which also identified the largest 1+2b=2a recombination block spanning the middle portion of env (Figure 3B, individuals p and q). While our results extend the existence of two subgroups across the entire SRLV genome, subgroup 2 has additional population sub-structuring (Figure 2, Figure 3B). Subgroup 2a may represent a somewhat stable locally circulating recombinant of subgroup 1 and 2b (Figure 3). The genomes identified as 2a were found exclusively in TMEM154 2.2 and 2.3 diplo-type animals (Table 1). Additionally, most strains with genomic regions indicative of dual infection were from TMEM154 susceptible 1.2 and 1.3 heterozygotes, i.e. animals with both an E and K at position 35. This suggests that animals that are E35K heterozygous due to TMEM154 1.2 and 1.3 diplo-types may facilitate recombination between subgroups 1 and 2.

Interestingly, two subtype specific functional variants were found in a region of the env gene variable region 4 (V4) which was recently identified to contain “signature patterns” associated with different clinical status in sheep and goats (Mendez et al., 2020). This region of V4 also contains targets of neutralizing antibodies and is predicted to play a role in virus entry (Skraban et al., 1999). Multiple amino acid changes were also observed in the N-terminus of env. None of the amino acids were predicted to change the env signal peptide cleavage site; however, it would be interesting to know if the five subgroup-specific amino acids affect post-translational modifications such as cleavage timing, folding, or glycosylation, phenomena documented to affect HIV fitness (Asmal et al., 2011; Snapp et al., 2017; Upadhyay et al., 2018). As more genomes are sequenced, and we learn more about the function of TMEM154 in the context of the virus lifecycle, it will be interesting to see which, if any, of these viral sequences are biologically responsible for TMEM154 associations.

**Data availability**

**Underlying data**

NCBI sequence accessions are provided in Table 1.

**Extended data**


This project contains the following extended data:

- Supplemental Table 1. Four hundred and thirteen diagnostic SNPs distinguishing SRLV A2 subgroups.
- Supplemental Table 2: Seventy-three genomic regions indicative of dual subgroup infection among 9 SRLV A2 genomes.
- Supplemental Table 3: Within-host amino acid variability (functional viral quasispecies) at 106 subgroup diagnostic loci. Colors according to subgroup 1 vs subgroup 2 recombination model (Figure 4) can be seen in the downloaded file.

Figshare: Supplemental Figure 1. [https://doi.org/10.6084/m9.figshare.13109990](https://doi.org/10.6084/m9.figshare.13109990) (Dickey & Workman, 2020b).

This file contains 23 SRLV subtype A2 genomes ‘painted’ with recipient genomic ‘chunks’ derived from populations of donor genomes. The five donor genome populations in the recombination model were determined using FineStructure.

Extended data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

**Author contributions**

Conceptualization (AMW, MLC, MPH); Data Curation (AMW, AMD, TPLS); Formal Analysis (AMW, AMD); Investigation (AMW, TPLS); Methodology (AMW, AMD, TPLS); Project Administration (AMW); Resources (AMW, MPH, MLC, AMD, TPLS); Software (AMD); Visualization (AMD, AMW); Writing (AMD, AMW); Editing (AMW, MLC, MPH, AMD, TPLS)

**Acknowledgments**

Sue Hauver, Brad Sharp, William Thompson, Jacky Carnahan, Kelsey McClure and Kristen Kuhn assisted with RNA library
preparation and sequencing. Stephanie Schmidt provided secretarial support. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity provider and employer.

References


Dickey A, Workman A: Supplemental Figure 1. figshare. Figure. 2020b. http://www.doi.org/10.6084/m9.figshare.13109909.v1


Shankarapappa R, Margolick JB, Gange SJ, et al.: Consistent viral evolutionary


Open Peer Review

Current Peer Review Status: ??

Version 1

Reviewer Report 28 April 2021

https://doi.org/10.5256/f1000research.30850.r83037

© 2021 Loncoman C. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Carlos Loncoman

Applied Biochemistry Laboratory, Institute of Pharmacology and Morphophysiology, Faculty of Veterinary Sciences, Universidad Austral de Chile, Valdivia, Chile

In this manuscript the authors assembled full length viral genomes and then analyzed specific segments in order to identify recombination patterns by using chromosome painting model.

General comments:

It's not clear in the introduction whether this chromosome painting model is used for recombination detection in any other virus? Perhaps any other RNA virus at all? I would highly appreciate if the authors could expand their introduction about this.

I would suggest the authors to include RDP software in methods and at least 1 figure in the results to see where the recombination events take place within the genome and compare these results with the chromosome painting one. Currently, these results are as not shown so it's not very clear whether these results are complementary.

My major concern with the recombination events that the authors describe in the manuscript are related with the differentiation between mutation and recombination. They mention in the introduction that RNA viruses have more mutations due to the low fidelity of the reverse transcriptase how did they differentiate these events from true recombination events?

Specific comments:

Abstract - Methods – The goal. Could the author explain the method rather than the goal?

In the first paragraph would be good to read 1 or 2 lines about the importance or impact of this pathogen over sheep population not just the disease that cause the virus. Perhaps include incidence or any statistics about the disease that cause the virus. This would help to put this research in context.
As the genome contains more than one non-coding long terminal repeat regions. I wonder how did the authors manage to sequence that specific part as I know they could be problematic to determine sequence identity. Are these regions inverted?

Why the authors did not amplify the virus previous to sequencing? Is there any cell culture that could be used for this purposes? Could this help with the four strains that had low quality long read sequence?

RDP software is mentioned in the discussion but not in the methods.

The split tree software has a tool to measure recombination called PHI test. Is there a reason why the authors did not include this in the manuscript?

What does PC stands for in Figure 2?

Is the work clearly and accurately presented and does it cite the current literature? Yes

Is the study design appropriate and is the work technically sound? Yes

Are sufficient details of methods and analysis provided to allow replication by others? Partly

If applicable, is the statistical analysis and its interpretation appropriate? I cannot comment. A qualified statistician is required.

Are all the source data underlying the results available to ensure full reproducibility? Partly

Are the conclusions drawn adequately supported by the results? Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Virology; recombination; herpesvirus

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 16 Jul 2021

Aaron Dickey, US Department of Agriculture, Agricultural Research Service, US Meat Animal Research Center, Clay Center, USA
General comments:

1. It’s not clear in the introduction whether this chromosome painting model is used for recombination detection in any other virus? Perhaps any other RNA virus at all? I would highly appreciate if the authors could expand their introduction about this.

Response: The recombination modeling program utilized in these analyses has been previously used to analyze recombination in Helicobacter pylori prophages (Yahara et al., 2019) and Herpes Simplex Virus (Forni et al., 2020). This information has been added to the methods. The Yahara et al. reference may be most similar in initial approach. Ours may be the first use of ChromoPainter version 2 ‘donor mode’ to extract subgroup-diagnostic variants in viruses while accounting for recombination.

2. I would suggest the authors to include RDP software in methods and at least 1 figure in the results to see where the recombination events take place within the genome and compare these results with the chromosome painting one. Currently, these results are as not shown so it’s not very clear whether these results are complementary.

Response: We agree with the reviewer that our use of RDP belongs in the methods. We have attempted to do so in a way that does not distract from the primary focus of the manuscript, which was to identify subgroup-diagnostic variants, rather than recombination events. RDP was only used in this work to confirm a single putative event: the possible introgression of a large chunk of subgroup 1 sequence into a subgroup 2b background to create subgroup 2a (black box in Figure 3b). This event (approximate size, genomic position, and ChromoPainter assigned subgroup of RDP identified putative parental strains) was confirmed by RDP. This information has been added to the results and the details of the RDP event are provided in Extended data, Supplemental Table 2(Dickey & Workman, 2020a).

3. My major concern with the recombination events that the authors describe in the manuscript are related with the differentiation between mutation and recombination. They mention in the introduction that RNA viruses have more mutations due to the low fidelity of the reverse transcriptase how did they differentiate these events from true recombination events?

Response: All recombination analysis programs claim to account for mutation as this is a fundamental issue. ChromoPainter initializes at the mutation rate of Li and Stephens 2003, but we estimated it in our analyses by using the -iM flag according to program developer recommendations. Interestingly, ChromoPainter’s mutation rate estimates are 3-fold higher for subgroup 2 genomes compared to subgroup 1 (Table 2). Population specific mutation rate estimates were obtained by averaging the estimated rates from the fineSTRUCTURE run among members of the assigned population and incorporated into our increasingly refined recombination models using the -m flag. We have added this information and the Li and Stephens reference to the manuscript. Our subgroup-specific-colored blocks in figures 3-5 are best interpreted as regions of sequence assigned by the recombination model to those subgroups, not as events-per-se. We believe regions alternating back-and-forth between subgroups over short genomic spans (e.g. strains labeled k-r in figures 3-5), are sometimes best interpreted as having arisen from multiple recombination events over time. We have attempted to minimize the use of the term ‘event’ in the manuscript in service of this paradigm.

Specific comments:

1. Abstract - Methods – The goal. Could the author explain the method rather than the goal? Response: The methods section now reads: Methods: Sequence variation within and among full-length assemblies of SRLV subtype A2 subgroups 1 and 2 was analyzed to identify genome-scale recombination patterns.

2. In the first paragraph would be good to read 1 or 2 lines about the importance or impact of this pathogen over sheep population not just the disease that cause the virus. Perhaps include incidence or any statistics about the disease that cause the virus. This would help to put this research in context. Response: This information has been added to the first paragraph of the introduction.

3. As the genome contains more than one non-coding long terminal repeat regions. I wonder how did the authors manage to sequence that specific part as I know they could be problematic to determine sequence identity. Are these regions inverted? Response: We used PacBio long-read technology to help resolve the ends of the viral genome. We report some genomes as ‘near-complete’ because we did not obtain the complete LTR sequence for all genomes.

4. Why did the authors not amplify the virus previous to sequencing? Is there any cell culture that could be used for this purposes? Could this help with the four strains that had low quality long read sequence? Response: It is true that the virus can be isolated in culture prior to sequencing; however, there were no commercially available cell lines for the isolation of SRLV at the time we initiated this project. While we could have developed appropriate cell lines, we chose not to isolate the virus first, as this can lead to the selection of a subpopulation of the infecting virus quasispecies. For these reasons, we sequenced the virus directly from the infected tissues.

5. RDP software is mentioned in the discussion but not in the methods. Response: Please see the response to General comment 2.

6. The split tree software has a tool to measure recombination called PHI test. Is there a reason why the authors did not include this in the manuscript? Response: SplitsTree5 does not have this analysis option, though we also remember it from SplitsTree4. We ran a PHI-test in RDP5 and now report it in the results.

7. What does PC stand for in Figure 2? Response: Principal Component (PC), is now incorporated into the legend of Figure 2.

Competing Interests: No competing interests were disclosed.

Reviewer Report 08 February 2021

https://doi.org/10.5256/f1000research.30850.r77216

© 2021 Olech M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Monika Olech
In this article, the authors focused mainly on the analysis of recombination of 23 SRLV strains belonging to subtype A2. In this paper a lot of different analyses were performed, but the results and figures are described very poorly. This makes the article very hard to read. For example on Fig. 1 I see only 15 isolates not 23. Lack is information on accession numbers or isolate names of sequences from GenBank present on the tree. Please indicate to which genotypes they belong. In Figure 2, there are three images that should be labeled as A, B and C and their description should be more detailed (in figure and in results). In the third picture I see three populations not six.

I don't understand why all 23 sequences were used in the recombination analysis. The authors write that only subgroups 2a, 2b and 1 was used in the recombination models but some genomes not belonged to these subgroups (fine structure Pop4, 5 and 6, Figure2).

In the discussion the results of the study are repeated.

There are no analyses regarding the occurrence of TMEM154E35K genotypes and susceptibility/resistance of sheep infected with and SRLV A2 subgroups 1 and 2. Did the research confirm previous studies described by Sider et al., 2013?

**Is the work clearly and accurately presented and does it cite the current literature?**
Partly

**Is the study design appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**
Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
I cannot comment. A qualified statistician is required.

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Genetic and antigenic characterization of Small Ruminant Lentiviruses (SRLV) in sheep, goats and wildlife ruminants (molecular analysis, phylogeny, cross species infection, genetic recombination)

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
1. In this article, the authors focused mainly on the analysis of recombination of 23 SRLV strains belonging to subtype A2. In this paper a lot of different analyses were performed, but the results and figures are described very poorly. This makes the article very hard to read. For example on Fig. 1 I see only 15 isolates not 23. Response: Figure 1 is information dense, particularly the very similar subgroup 1 genomes denoted by the lavender-colored oval. However, we agree it is beneficial to list all genomes, so they are now listed in clockwise order in the new Supplemental Table 1.

2. Lacking is information on accession numbers or isolate names of sequences from GenBank present on the tree. Response: An added benefit to the new supplemental table (see previous comment) is space to put the GenBank accession numbers of all strains in Figure 1. Genbank accession number, isolate name, county of origin, and virus genotype are now listed in Supplemental Table 1.

3. Please indicate to which genotypes they belong. Response: This information is made clearer in the figure by curly brackets, additional annotations, and clarification in the legend.

4. In Figure 2, there are three images that should be labeled as A, B and C and their description should be more detailed (in figure and in results). In the third picture I see three populations not six. Response: The figures sections have been relabeled A, B and C and additional details are provided in the results. We are not certain but think the reviewer may be referring to 3 clusters of strains, not six, since principal components 2 and 3 (PC2 and PC3) do not strongly discriminate subgroups 1 and 2b. We do not find this problematic since the first principal component strongly discriminates these two populations.

5. I don’t understand why all 23 sequences were used in the recombination analysis. The authors write that only subgroups 2a, 2b and 1 was used in the recombination models but some genomes not belonged to these subgroups (Fine structure Pop4, 5 and 6, Figure2). Response: Our initial recombination donor model had five donor populations and all 23 genomes as possible recipients (Supplemental Figure 1). fineSTRUCTURE population 6 was excluded from this analysis on the grounds that singletons could not comprise a population for this type of analysis. The result suggests fineSTRUCTURE populations 4 and 5 are probably not populations either, justifying our continued refinement of recombination models to those with fewer donor populations, which better characterize the subgroups for which there is sufficient evidence, and recombination patterns among them. Our explanation of this is that fineSTRUCTURE assigned ‘population’ status to groups of genomes that would be better described as complex recombinants of already defined subgroups, rather than true populations. Over-splitting of these types of datasets into too many ‘populations’ is a known problem. See: Falush D, van Dorp L, Lawson D, 2016 A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots. bioRxiv 066431; 10.1101/066431. While, fewer potential donor genomes were used in the refined recombination models (Figures 3 and 4), all genomes were used as recipients,
allowing us to characterize recombination.

6. In the discussion the results of the study are repeated. Response: We believed this was beneficial to organizing points of discussion.

7. There are no analyses regarding the occurrence of TMEM154E35K genotypes and susceptibility/resistance of sheep infected with and SRLV A2 subgroups 1 and 2. Did the research confirm previous studies described by Sider et al., 2013? Response: We generated complete genome sequences from a subset of genomes characterized in the Sider et al., 2013 paper. The goal was not to confirm or dispute the association, but rather, to generate complete genomes and develop a nucleotide polymorphism typing system for the two major subgroups while accounting for recombination. More complete SRLV sequences are required to identify which viral elements correlate with host adaptation.

**Competing Interests:** No competing interests were disclosed.