bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

# Pooled CRISPR Inverse PCR sequencing (PCIP-seq): simultaneous sequencing of retroviral insertion points and the associated provirus in thousands of cells with long reads

Maria Artesi[1,2,7], Vincent Hahaut[1,2,7], Fereshteh Ashrafi[1,3], Ambroise Marçais[4], Olivier Hermine[4], Philip Griebel[5], Natasa Arsic[5], Frank van der Meer[6], Arsène Burny[2], Dominique Bron[2], Carole Charlier[1], Michel Georges[1], Anne Van den Broeke[1,2,8,*], Keith Durkin[1,2,8,*]

[1]Unit of Animal Genomics, GIGA-R, Université de Liège (ULiège), Avenue de l'Hôpital 11, B34, Liège 4000, Belgium. [2]Laboratory of Experimental Hematology, Institut Jules Bordet, Université Libre de Bruxelles (ULB), Boulevard de Waterloo 121, Brussels 1000, Belgium. [3]Department of Animal Science, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran [4]Service d'hématologie, Hôpital Universitaire Necker, Université René Descartes, Assistance Publique Hôpitaux de Paris, Paris, France. [5]Vaccine and Infectious Disease Organization, VIDO-Intervac, University of Saskatchewan, 120 Veterinary Road, Saskatoon, Canada S7N 5E3, [6]Faculty of Veterinary Medicine: Ecosystem and Public Health, Calgary, AB, Canada, [7]These authors contributed equally: Maria Artesi, Vincent Hahaut, [8]These authors jointly supervised the work: Anne Van den Broeke, Keith Durkin, *corresponding authors, email: anne.vandenbroeke@bordet.be, kdurkin@uliege.be

## Abstract

Retroviral infections create a large population of cells, each defined by a unique proviral insertion site. Methods based on short-read high throughput sequencing can identify thousands of insertion sites, but the proviruses within remain unobserved. We have developed Pooled CRISPR Inverse PCR sequencing (PCIP-seq), a method that leverages long reads on the Oxford Nanopore MinION platform to sequence the insertion site and its associated provirus. We have applied the technique to three exogenous retroviruses, HTLV-1, HIV-1 and BLV, as well as endogenous retroviruses in both cattle and sheep. The long reads of PCIP-seq improved the accuracy of insertion site identification in repetitive regions of the genome. The high efficiency of the method facilitated the identification of tens of thousands of insertion sites in a single sample. We observed thousands of SNPs and dozens of structural variants within proviruses and uncovered evidence of viral hypermutation, recombination and recurrent selection.

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## Introduction

The integration of viral DNA into the host genome is a defining feature of the retroviral life cycle, irreversibly linking provirus and cell. This intimate association facilitates viral persistence and replication in somatic cells, and with integration into germ cells bequeaths the provirus to subsequent generations. Considerable effort has been expended to understand patterns of proviral integration, both from a basic virology stand point, and due to the use of retroviral vectors in gene therapy[1]. The application of next generation sequencing (NGS) over the last ~10 years has had a dramatic impact on our ability to explore the landscape of retroviral integration for both exogenous and endogenous retroviruses. Methods based on ligation mediated PCR and Illumina sequencing have facilitated the identification of hundreds of thousands of insertion sites in exogenous viruses such as Human T-cell leukemia virus-1 (HTLV-1)[2] and Human immunodeficiency virus (HIV-1)[3]. These techniques have shown that in HTLV-1[2], Bovine Leukemia Virus (BLV)[4], Avian Leukosis Virus (ALV)[5] and HIV-1[6] integration sites are not random, most likely reflecting clonal selection. In HIV-1 it has also become apparent that provirus integration can drive clonal expansion[3] and contribute to the HIV-1 reservoir, placing a major road block in the way of a complete cure.

Current methods based on short-read sequencing identify the insertion point but the provirus its-self is unexplored. Whether variation in the provirus influences the fate of the clone remains difficult to investigate. Using long range PCR it has been shown that proviruses in HTLV-1 induced Adult T-cell leukemia (ATL) are frequently (~45%) defective[7], although the abundance of defective proviruses within asymptomatic HTLV-1 carriers has not been systematically investigated. Recently, there has been a concerted effort to better understand the structure of HIV-1 proviruses in the latent reservoir. Methods such as Full-Length Individual Proviral Sequencing (FLIPS) have been developed to identify functional proviruses[8], however these methods do not provide information on the provirus integration site, removing the possibility of tracking clone abundance and persistence over time.

Retroviruses are primarily associated with the diseases they provoke through the infection of somatic cells. Over the course of evolutionary time they have also played a major role in shaping the genome. Retroviral invasion of the germ line has occurred multiple times, resulting in the remarkable fact that endogenous retrovirus (ERV)-like elements comprise a larger proportion of the human genome (8%) than protein coding sequences (~1.5%)[9]. With the availability of multiple vertebrate genome assemblies, much of the focus has been on comparison of ERVs between species. However, single genomes represent a fraction of the variation within a species, prompting some to take a population approach to investigate ERV–

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

host genome variation[10]. While capable of identifying polymorphic ERVs in the population, approaches relying on conventional paired-end libraries and short reads cannot capture the sequence of the provirus beyond the first few hundred bases of the proviral long terminal repeat (LTR), leaving the variation within uncharted.

The application of NGS as well as Sanger sequencing before, has had a large impact on our understanding of both exogenous and endogenous proviruses. The development of long-read sequencing, linked-read technologies and associated computational tools[11] have the potential to explore questions inaccessible to short reads. Groups investigating Long interspersed nuclear elements-1 (LINE-1) insertions[12] and the koala retrovirus, KoRV[13] have highlighted this potential and described techniques utilizing the Oxford Nanopore and PacBio platforms, to investigate insertion sites and retroelement structure.

To more fully exploit the potential of long reads we developed Pooled CRISPR Inverse PCR sequencing (PCIP-Seq), a method that leverages selective cleavage of circularized DNA fragments carrying proviral DNA with a pool of CRISPR guide RNAs, followed by inverse long-range PCR and multiplexed sequencing on the Oxford Nanopore MinION platform. Using this approach, we can now simultaneously identify the integration site and track clone abundance while also sequencing the provirus inserted at that position. We have successfully applied the technique to the retroviruses HTLV-1, HIV-1 and BLV as well as endogenous retroviruses in cattle and sheep.

## Results

### Overview of PCIP-seq (Pooled CRISPR Inverse PCR-sequencing)

The genome size of the retroviruses targeted ranged from 6.8 to 9.7kb, therefore we chose to shear the DNA to ~8kb in length. In most cases this creates two fragments for each provirus, one containing the 5’ end with host DNA upstream of the insertion site and the second with the 3’ end and downstream host DNA. Depending on the shear site the amount of host and proviral DNA in each fragment will vary (Fig. 1a). To facilitate identification of the provirus insertion site we carry out intramolecular ligation, followed by digestion of the remaining linear DNA. To selectively linearize the circular DNA containing proviral sequences, regions adjacent to the 5’ and 3’ LTRs in the provirus are targeted for CRISPR mediated cleavage. We sought a balance between ensuring that the majority of the reads contained part of the flanking DNA (for clone identification) while also generating sufficient reads extending into the midpoint of the provirus. We found that using a pool of CRISPR guides for each region increased the efficiency and by multiplexing the guide pools and PCR primers

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for the 5' and 3' ends we could generate coverage for the majority of the provirus in a single reaction (Fig. 1b). The multiplexed pool of guides and primers leaves coverage gaps in the regions flanked by the primers. To address these coverage gaps we designed a second set of guides and primers. Following separate CRISPR cleavage and PCR amplification the products of these two sets of guides and primers were combined for sequencing (Fig. 1c). This approach ensured that the complete provirus was sequenced (Fig. 1d).

**Identifying genomic insertions and internal variants in human retroviruses**

Adult T-cell leukemia (ATL) is an aggressive cancer induced by HTLV-1. It is generally characterized by the presence of a single dominant malignant clone, identifiable by a unique proviral integration site. We and others have developed methods based on ligation mediated PCR and Illumina sequencing to simultaneously identify integration sites and determine the abundance of the corresponding clones. We initially applied PCIP-seq to two HTLV-1 induced cases of ATL, both previously analyzed with our Illumina based method (ATL2[4] & ATL100[14]). In ATL100 both methods identify a single dominant clone, with >95% of the reads mapping to a single insertion site on chr18 (Fig. 2a, 2b & Table1). Using the integration site information, we extracted the PCIP-seq hybrid reads spanning the provirus/host insertion site, uncovering a ~3,600bp deletion within the provirus (Fig. 2c).

In the case of ATL2, PCIP-seq showed three major proviruses located on chr5, chr16 and chr1, each responsible for 33.9%, 33.2% and 31.2% of the HTLV-1/host hybrid reads respectively. We had previously established that these three proviruses are in a single clone via examination of the T-cell receptor gene rearrangement[4]. However, it is interesting to note that this was not initially obvious using our Illumina based method as the proviral insertion site on chr1 falls within a repetitive element (LTR) causing many of the reads to map to multiple regions in the genome. If multi mapping reads are filtered out, the chr1 insertion site accounted for 2.6% of the remaining reads, while retaining multi mapping produces values closer to reality (25.9%). In contrast the long reads from PCIP-seq allow unambiguous mapping and closely matched the expected ~33% for each insertion site (Fig. 2d), highlighting the advantage long reads have in repetitive regions. Looking at the three proviruses, proviral reads revealed all to be full length. Three de novo mutations were observed in one provirus and a single de novo mutation was identified in the second (Fig. 2e).

As a proof of concept, we also applied the technique to U1[15], a HIV-1 cell line containing replication competent proviruses, that shows evidence of ongoing viral replication[16]. As

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

expected, PCIP-seq found the major insertion sites on chr2 and chrX (accounting for 56% & 42% of the hybrid reads respectively). It has been reported that both proviruses have defective Tat function[17] and we observed the mutation disrupting the ATG initiation codon in the chr2 provirus and the H-to-L mutation at amino acid 13 in the chrX provirus (Supplementary Fig. 1a). In addition to the two major proviruses we identified an additional ~1,200 low abundance insertion sites (Table 1). We manually inspected the proviruses from the ~50 most abundant insertion sites. Of these, an insertion site on chr19 (0.3%) was reported by Symons et al 2017[16]. This provirus as well as a second on chr7 were found to be the products of recombination between the major chrX and chr2 proviruses (Supplementary Fig. 1b). Additionally, two proviruses showed high levels of G-to-A mutations, indicating hypermutation by the APOBEC3G protein, a component of the innate anti-viral immune response[18] (Fig. 3 & Supplementary Fig. 2). Taken together these results suggest PCIP-seq as a useful tool for analyzing the HIV-1 reservoir.

**Insertion sites identified in samples with multiple clones of low abundance**

The samples utilized above represent a best-case scenario, with ~100% of cells infected and a small number of major clones. However, samples like this are the exception rather than the rule in most infected individuals. To be broadly applicable, PCIP-seq should be able to identify thousands of insertion sites in samples with large numbers of low abundance clones, where each clone is defined via a unique proviral insertion site. It should also identify insertion sites when only a small fraction of the cells carries the provirus. To test these potential limitations, we applied PCIP-seq to four samples from BLV infected sheep (experimental infection[19]) and three cattle (natural infection) to explore its performance on polyclonal and low proviral load (PVL) samples and compared PCIP-seq to our previously published Illumina method[4]. PCIP-seq revealed all samples to be highly polyclonal (Supplementary Fig. 3 and Table 1) with the number of insertion sites identified varying from 233 in the bovine sample 560 (1µg template, PVL 0.644%) to 21,579 in bovine sample 1053 (6µg template, PVL 23.5%). In general, PCIP-seq identified more insertion sites, using less input DNA than our Illumina based method (Supplementary Table 1). Comparison of the results showed a significant overlap between the two methods. When we consider insertion sites supported by more than three reads in both methods (larger clones, more likely to be present in both samples), in the majority of cases >70% of the insertion sites identified in the Illumina data were also observed via PCIP-seq (Supplementary Table 1). These results show the utility of PCIP-seq for insertion site

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

identification, especially considering the advantages long reads have in repetitive regions of the genome.

**Identifying SNPs in BLV proviruses**

Proviruses with more than 10 supporting reads were examined for SNPs with LoFreq[20]. For the four sheep samples, the variants were called relative to the pBLV344 provirus (used to infect the animals). For the bovine samples 1439 and 1053 custom consensus BLV sequences were generated for each and the variants were called in relation to the appropriate reference (SNPs were not called in 560). Across all the samples 20,283 proviruses were examined, 3,085 SNPs were called and 2,565 (13%) of the proviruses carried one or more SNPs (Supplementary Table 2). We validated 10 BLV SNPs in the ovine samples and 15 in the bovine via clone specific long-range PCR and Illumina sequencing (Supplementary Fig. 4). For Ovine 221, which was sequenced twice with an ~2-year interval, we identified six instances where the same SNP and provirus were observed at both time points. We noted a small number of positions in the BLV provirus prone to erroneous SNP calls. By comparing allele frequencies from bulk Illumina and Nanopore data these problematic positions could be identified (Supplementary Fig. 5a).

Within each individual, between ~30% and ~49% of the SNPs were found in multiple proviruses. Generally, SNPs found at the same position in multiple proviruses were concentrated in a single individual, indicating their presence in a founder provirus or via a mutation in the very early rounds of viral replication (Supplementary Fig. 5b). Alternatively, a variant may also rise in frequency due to increased fitness of clones carrying a mutation in that position. In this instance, we would expect to see the same position mutated in multiple individuals. One potential example is found in the first base of codon 303 (position 8155) of the viral protein Tax, a potent viral transactivator, stimulator of cellular proliferation and highly immunogenic[21]. A variant was observed at this position in multiple proviruses (14, 18 and 2 respectively) in the ovine samples 233, 221 (022016) & 220, in one provirus from bovine 1053 and in two for bovine 1439 (Fig. 4). The majority of the variants observed were G-to-A transitions (results in E-to-K amino acid change), however we also observed G-to-T (E-to-STOP) and G-to-C (E-to-Q) transversions. It has been previously shown that the G-to-A mutation abolishes the Tax proteins transactivator activity[21,22]. The repeated selection of variants in this specific position suggests that they reduce viral protein recognition by the immune system, while preserving the Tax proteins other proliferative properties.

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Patterns of provirus-wide hypermutation (G-to-A) seen within HIV-1 were not observed in BLV. However, three proviruses (two from sheep 233 and one in bovine 1053) showed seven or more A-to-G transitions, confined to a ~70bp window in the first half of the U3 portion of the 3'LTR. The pattern of mutation, as well as their location in the provirus suggests the action of RNA adenosine deaminases 1 (ADAR1)[23,24].

**PCIP-seq identifies BLV structural variants in multiple clones**

The same set of proviruses (with more than 10 supporting reads) used for calling SNPs were also examined for structural variants (SVs) using a custom script and visualization of proviruses (see methods). Between the sheep and bovine samples, we identified 66 deletions and 3 tandem duplications, with sizes ranging from 15bp to 4,152bp, with a median of 113bp (Supplementary Table 3). We validated 14 of these via clone specific PCR (Supplementary Fig. 7). As seen in Fig. 5 SVs were found throughout the majority of the provirus, encompassing the highly expressed microRNAs[25] as well as the second exon of the constitutively expressed antisense transcript *AS1*[26]. Only two small regions at the 3' end lacked any SVs. More proviruses will need to be examined to see if this pattern holds, but these results again suggest the importance of the 3'LTR and its previously reported interactions with adjacent host genes[4].

**Identifying full-length and polymorphic endogenous retroviruses in cattle and sheep**

ERVs in the genome are generally present as full length, complete provirus, or more commonly as solo-LTRs, the products of non-allelic recombination[27]. At the current time conventional short read sequencing, using targeted or whole genome approaches, cannot distinguish between the two classes. Examining full length ERVs would provide a more complete picture of ERV variation, while also revealing which elements can produce de novo ERV insertions. As PCIP-seq targets inside the provirus we can preferentially amplify full length ERVs, opening this type of ERV to study in larger numbers of individuals. As a proof of concept we targeted the class II bovine endogenous retrovirus BERVK2, known to be transcribed in the bovine placenta[28]. We applied the technique to three cattle, one (10201e6) was a Holstein suffering from cholesterol deficiency, an autosomal recessive genetic defect recently ascribed to the insertion of a 1.3kb LTR in the *APOB* gene[29]. PCIP-seq clearly identified the *APOB* ERV insertion in 10201e6 and in contrast to previous reports[29] shows it to be a full-length element (Supplementary Fig. 8). We identified a total of 67 ERVs, with 8 present in all three samples (Supplementary Table 4). We validated three ERVs via long

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

range PCR and Illumina sequencing (Supplementary Fig. 9). We did not find any ERVs with an identical sequence to the *APOB* ERV, although the ERV BTA3_115.3 has an identical LTR sequence, highlighting that the sequence of the LTR cannot be used to infer the complete sequence of the ERV (Supplementary Fig. 10).

We also adapted PCIP-seq to amplify the Ovine endogenous retrovirus Jaagsiekte sheep retrovirus (enJSRV), a model for retrovirus-host co-evolution[30]. Using two sheep (220 & 221) as template we identified a total of 48 enJSRV proviruses, (33 in 220 and 38 in 221, with 22 common to both) and of these ~54% were full length (Supplementary Table 5). We validated seven proviruses via long-range PCR and Illumina sequencing (Supplementary Fig. 11).

## Discussion

In the present report we describe how PCIP-seq can be utilized to identify insertion sites while also sequencing the associated provirus and confirm this methodology is effective with a number of different retroviruses. For insertion site identification, the method was capable of identifying more than ten thousand BLV insertion sites in a single sample, using ~4µg of template DNA. Even in samples with a PVL of 0.66%, it was possible to identify hundreds of insertion sites with only 1µg of DNA as template. The improved performance of PCIP-seq in repetitive regions further highlights its utility, strictly from the standpoint of insertion site identification. In addition to its application in research, high throughput sequencing of retrovirus insertion sites has shown promise as a clinical tool to monitor ATL progression[14]. Illumina based techniques require access to a number of capital-intensive instruments. In contrast PCIP-seq libraries can be generated, sequenced and analyzed with the basics found in most molecular biology labs, additionally, preliminary results are available just minutes after sequencing begins[31]. As a consequence, the method may have use in a clinical context to track clonal evolutions in HTLV-1 infected individuals, especially as the majority of HTLV-1 infected individuals live in regions of the world with poor biomedical infrastructure[32].

The inability to easily link variation in the provirus to a specific insertion site is an especially pressing problem for the HIV-1 reservoir. Only a small fraction of proviruses (2.4%) in the reservoir are intact, yet these are more than sufficient for the disease to rebound if antiretroviral therapy is removed[33]. As strategies are developed to target these intact proviruses it will be essential to distinguish between the intact and defective proviruses[33]. Using PCIP-seq in the HIV-1 cell line U1 we identified thousands of minor insertion sites, readily finding examples of recombination and hypermutation. This approach could be

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

applied equally well to monitoring the abundance and persistence of intact HIV-1 proviruses in patients.

When analyzing SNPs from BLV two results in particular stood out. Firstly, the presence of the recurrent mutations at the first base of codon 303 in the viral protein Tax, a central player in the biology of both HTLV-1[32] and BLV[34]. It has previously been reported that this mutation causes an E-to-K amino acid substitution which ablates the transactivator activity of the Tax protein[21]. Collectively, these observations suggest this mutation confers an advantage to clones carrying it, possibly contributing to immune evasion, while retaining Tax protein functions that contribute to clonal expansion. However, there is a cost to the virus as this mutation prevents infection of new cells due to the loss of Tax mediated transactivation of the proviral 5'LTR making it an evolutionary dead end. The second striking observation is ADAR1 mediated hypermutation of a ~70bp window in the 3'LTR. ADAR1 hypermutation has been observed in a number of viruses[23], including the close BLV relatives HTLV-2 and simian T-cell leukemia virus type 3 (STLV-3)[35]. Given its propensity to mutate double stranded RNA it is somewhat surprising that ADAR1 has been co-opted by retroviruses to modulate the antiviral innate immune response[23]. Given the small number of hypermutated proviruses observed, it appears that this interplay does not result in large numbers of mutated proviruses in vivo, although it will be interesting to see it this holds for different retroviruses and at different time points during infection.

In the current study we focused our analysis on retroviruses and ERVs. However, this methodology is potentially applicable to a number of different targets. It is estimated that 10-15% of all cancers have a viral cause, HTLV-1 only represents a small fraction of this number, with human papillomaviruses (HPV) and hepatitis B virus (HBV) playing a much bigger role[36]. These viruses are often found integrated into the genome of the cancers they provoke[36]. In contrast to retroviruses, the viral sequence immediately adjacent to the integration site is not consistent, reducing the utility of ligation mediated PCR approaches. As a consequence, probe based capture methods are required to obtain a genome-wide picture of integration sites in these viruses[37,38]. Using PCIP-seq guides and primers, multiplexed at different points in the viral genome, combined with long-reads, it will be possible to identify the viral insertion site while also sequencing the virus. This could open many more viral integration sites and viral sequences to interrogation, while also exploring precancerous tissue for rare viral integrations. Other potential applications include determining the insertion sites and integrity of retroviral vectors[39], detecting transgenes in genetically modified organisms or identifying on target CRISPR–Cas9 mediated structural rearrangement that could be missed by

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

conventional long range PCR[40]. We envision that in addition to the potential applications outlined above many other novel targets/questions could be addressed using this method.

## References


1. Bushman, F. *et al.* Genome-wide analysis of retroviral DNA integration. *Nat Rev Micro* **3,** 848–858 (2005).
2. Gillet, N. A. *et al.* The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. *Blood* **117,** 3113–3122 (2011).
3. Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. & Bushman, F. D. HIV integration site selection: Analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Research* **17,** 1186–1194 (2007).
4. Rosewick, N. *et al.* Cis-perturbation of cancer drivers by the HTLV-1/BLV proviruses is an early determinant of leukemogenesis. *Nature Communications* **8,** 15264 (2017).
5. Malhotra, S. *et al.* Selection for avian leukosis virus integration sites determines the clonal progression of B-cell lymphomas. *PLoS Pathog* **13,** e1006708–25 (2017).
6. Singh, P. K. *et al.* LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes & Development* **29,** 2287–2297 (2015).
7. Miyazaki, M. *et al.* Preferential selection of human T-cell leukemia virus type 1 provirus lacking the 5' long terminal repeat during oncogenesis. *Journal of Virology* **81,** 5714–5723 (2007).
8. Hiener, B. *et al.* Identification of Genetically Intact HIV-1 Proviruses in Specific CD4+ T Cells from Effectively Treated Participants. *CellReports* **21,** 813–822 (2017).
9. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
10. Rivas-Carrillo, S. D., Pettersson, M. E., Rubin, C.-J. & Jern, P. Whole-genome comparison of endogenous retrovirus segregation across wild and domestic host species populations. *PNAS* **115,** 11012–11017 (2018).
11. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nature Reviews Genetics* **17,** 1–18 (2018).
12. Pradhan, B. *et al.* Detection of subclonal L1 transductions in colorectal cancer by long-distance inverse-PCR and Nanopore sequencing. *Scientific Reports* **7,** 1–12 (2017).
13. Löber, U. *et al.* Degradation and remobilization of endogenous retroviruses by recombination during the earliest stages of a germ-line invasion. *Proceedings of the National Academy of Sciences* **5,** 201807598–15 (2018).
14. Artesi, M. *et al.* Monitoring molecular response in adult T-cell leukemia by high-throughput sequencing analysis of HTLV-1 clonality. *Leukemia* **31,** 2532–2535 (2017).
15. Folks, T. M., Justement, J., Kinter, A., Dinarello, C. A. & Fauci, A. S. Cytokine-induced expression of HIV-1 in a chronically infected promonocyte cell line. *Science* **238,** 800–802 (1987).
16. Symons, J. *et al.* HIV integration sites in latently infected cell lines: evidence of ongoing replication. *Retrovirology* **14,** 1–11 (2017).
17. Emiliani, S. *et al.* Mutations in the tat Gene Are Responsible for Human Immunodeficiency Virus Type 1 Postintegration Latency in the U1 Cell Line. *Journal of Virology* **72,** 1666–1670 (1998).

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

18. Armitage, A. E. *et al.* APOBEC3G-Induced Hypermutation of Human Immunodeficiency Virus Type-1 Is Typically a Discrete 'All or Nothing' Phenomenon. *PLoS Genet* **8,** e1002550–12 (2012).
19. Willems, L. *et al.* In vivo infection of sheep by bovine leukemia virus mutants. *Journal of Virology* **67,** 4078–4085 (1993).
20. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Research* **40,** 11189–11201 (2012).
21. Van den Broeke, A. *et al.* In vivo rescue of a silent tax-deficient bovine leukemia virus from a tumor-derived ovine B-cell line by recombination with a retrovirally transduced wild-type tax gene. *Journal of Virology* **73,** 1054–1065 (1999).
22. Merimi, M. *et al.* Complete suppression of viral gene expression is associated with the onset and progression of lymphoid malignancy: observations in Bovine Leukemia Virus-infected sheep. *Retrovirology* **4,** 51 (2007).
23. Samuel, C. E. Adenosine deaminases acting on RNA (ADARs) are both antiviral and proviral. *Virology* **411,** 180–193 (2011).
24. Cachat, A. *et al.* ADAR1 enhances HTLV-1 and HTLV-2 replication through inhibition of PKR activity. *Retrovirology* **11,** 7415–15 (2014).
25. Rosewick, N. *et al.* Deep sequencing reveals abundant noncanonical retroviral microRNAs in B-cell leukemia/lymphoma. *Proceedings of the National Academy of Sciences* **110,** 2306–2311 (2013).
26. Durkin, K. *et al.* Characterization of novel Bovine Leukemia Virus (BLV) antisense transcripts by deep sequencing reveals constitutive expression in tumors and transcriptional interaction with viral microRNAs. *Retrovirology* **13,** 1–16 (2016).
27. Gemmell, P., Hein, J. & Katzourakis, A. Phylogenetic Analysis Reveals That ERVs 'Die Young' but HERV-H Is Unusually Conserved. *PLoS Comp Biol* **12,** e1004964 (2016).
28. Cornelis, G. *et al.* Captured retroviral envelope syncytin gene associated with the unique placental structure of higher ruminants. *Proceedings of the National Academy of Sciences* **110,** E828–E837 (2013).
29. Menzi, F. *et al.* A transposable element insertion in APOB causes cholesterol deficiency in Holstein cattle. *Animal Genetics* **47,** 253–257 (2016).
30. Arnaud, F. *et al.* A Paradigm for Virus–Host Coevolution: Sequential Counter-Adaptations between Endogenous and Exogenous Retroviruses. *PLoS Pathog* **3,** e170–14 (2007).
31. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530,** 228–232 (2016).
32. Bangham, C. R. M. Human T Cell Leukemia Virus Type 1: Persistence and Pathogenesis. *Annu. Rev. Immunol.* **36,** annurev–immunol–042617–053222–29 (2017).
33. Bruner, K. M. *et al.* A quantitative approach for measuring the reservoir of latent HIV-1 proviruses. *Nature* 1–19 (2019). doi:10.1038/s41586-019-0898-8
34. Gillet, N. *et al.* Mechanisms of leukemogenesis induced by bovine leukemia virus: prospects for novel anti-retroviral therapies in human. *Retrovirology* **4,** 18 (2007).
35. Ko, N. L., Birlouez, E., Wain-Hobson, S., Mahieux, R. & Vartanian, J. P. Hyperediting of human T-cell leukemia virus type 2 and simian T-cell leukemia virus type 3 by the dsRNA adenosine deaminase ADAR-1. *Journal of General Virology* **93,** 2646–2651 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

36. Tang, K.-W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M. & Larsson, E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nature Communications* **4,** 2513 (2013).
37. Hu, Z. *et al.* Genome-wide profiling of HPV integration in cervical cancer identifies clustered genomic hot spots and a potential microhomology-mediated integration mechanism. *Nat Genet* **47,** 158–163 (2015).
38. Zhao, L.-H. *et al.* Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. *Nature Communications* **7,** 12992 (2016).
39. Goodwin, L. O. *et al.* Large-scale discovery of mouse transgenic integration sites reveals frequent structural variation and insertional mutagenesis. *Genome Research* gr.233866.117 (2019). doi:10.1101/gr.233866.117
40. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nature Biotechnology* 1–10 (2018). doi:10.1038/nbt.4192

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

## Methods

### Samples

Both the BLV infected sheep[4] and HTLV-1 samples[4,14] have been previously described. Briefly, the sheep were infected with the molecular clone pBLV344[19], following the experimental procedures approved by the University of Saskatchewan Animal Care Committee based on the Canadian Council on Animal Care Guidelines (Protocol #19940212). The HTLV-1 samples[4,14] were obtained with informed consent following the institutional review board-approved protocol at the Necker Hospital, University of Paris, France, in accordance with the Declaration of Helsinki. The BLV bovine samples were natural infections, obtained from commercially kept adult dairy cows in Alberta, Canada. Sampling was approved by VSACC (Veterinary Sciences Animal care Committee) of the University of Calgary: protocol number: AC15-0159. The bovine 571 used for ERV identification was collected as part of this cohort. The two sheep samples used for Jaagsiekte sheep retrovirus (enJSRV) identification were the BLV infected ovine samples (220 & 221 (032014)), with a PVL of 3.8 and 16% respectively. PBMCs were isolated using standard Ficoll-Hypaque separation. The DNA for the bovine Mannequin was extracted from sperm, while the DNA for bovine 10201e6 was extracted from whole blood using standard procedures. The HIV-1 U1 cell line DNA was provided by Dr. Carine Van Lint, IBMM, Gosselies, Belgium. No statistical test was used to determine adequate sample size and the study did not use blinding.

### PCIP-seq

DNA isolation was carried out using the Qiagen AllPrep DNA/RNA/miRNA kit. High molecular weight DNA was sheared to ~8kb using Covaris g-tubesTM (Woburn, MA) or a Megaruptor (Diagenode), followed by end-repair using the NEBNext EndRepair Module (New England Biolabs). Intramolecular circularization was achieved by overnight incubation at 16$^{o}$C with T4 DNA Ligase. Remaining linear DNA was removed with Plasmid-Safe-ATP-Dependent DNAse (Epicentre, Madison WI). Guide RNAs were designed using chopchop (http://chopchop.cbu.uib.no/index.php). The EnGen™ sgRNA Template Oligo Designer (http://nebiocalculator.neb.com/#!/sgrna) provided the final oligo sequence. Oligos were synthesized by Integrated DNA Technologies (IDT). Oligos were pooled and guide RNAs synthesized with the EnGen sgRNA Synthesis kit, S. pyogenes (New England Biolabs). Selective linearization reactions were performed with the Cas-9 nuclease, S. pyogenes (New England Biolabs). PCR primers flanking the cut sites were designed using primer3 (http://bioinfo.ut.ee/primer3/). Primers were tailed to facilitate the addition of Oxford Nanopore

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

indexes in a subsequent PCR reaction. The linearized fragments were PCR amplified with LongAmp Taq DNA Polymerase (New England Biolabs) and a second PCR added the appropriate Oxford Nanopore index. PCR products were visualized on a 1% agarose gel and quantified on a Nanodrop spectrophotometer. Indexed PCR products were multiplexed and Oxford Nanopore libraries prepared with either the Ligation Sequencing Kit 1D (SQK-LSK108) or 1D^2 Sequencing Kit (SQK-LSK308). The resulting libraries were sequenced on Oxford Nanopore MinION R9.4 or R9.5 flow cells respectively and basecalled using albacore 2.3.1. Only the 1D reads from both flow cell versions were used.

**Identification of proviral integrations sites in PCIP-seq**

Reads were mapped with Minimap2[41] to the host genome with the proviral genome as a separate chromosome. In-house R-scripts were used to identify integration sites **(**IS). Briefly, chimeric reads that partially mapped to at least one extremity of the proviral genome were used to extract virus-host junctions and shear sites. Junctions within a 200bp window were clustered together to form an “IS cluster”, compensating for sequencing/mapping errors. The IS retained corresponded to the position supported by the highest number of virus-host junctions in each IS cluster. Clone abundance was estimated based on the number of reads supporting each IS cluster, reads with the same shear site were considered PCR duplicates. A detailed outline of the workflow is available on Github: https://github.com/GIGA-AnimalGenomics-BLV/PCIP-seq.

**Identification of proviral integration sites (Illumina) and measure of proviral load (PVL)**

Integration sites and PVL were determined as previously described in Rosewick et al 2017[4] and Artesi et al 2017[14].

**Variant Calling**

After PCR duplicate removal, IS supported by more than 10 reads were retained for further processing. SNPs were identified using LoFreq[20] with default parameters, only SNP with an allele frequency of >0.6 in the provirus associated with the insertion site were considered. Deletions were called with a in house R-scripts. Briefly, samtools pileup[42] was used to calculate/compute coverage and deletions at base resolution. We used the changepoint detection algorithm PELT[43] to identify genomic windows showing an abrupt change in coverage. Windows that showed at least a 4-fold increase in the frequency of deletions

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(absence of a nucleotide for that position within a read) were flagged as deletions and visually confirmed in IGV[44].

**HIV-1 proviral sequences**

Sequences of the two major proviruses integrated in chr2 and chrX of the U1 cell line were generated by initially mapping the reads from both platforms to the HIV-1 provirus, isolate NY5 (GenBank: M38431.1), where the 5'LTR sequence is appended to the end of the sequence to produce a full-length HIV-1 proviral genome reference. The sequence was then manually curated to produce the sequence for each provirus. To check for recombination, reads of selected clones were mapped to the sequence from the chrX provirus and the patterns of SNPs examined to determine if the variants matched the chrX or chr2 proviruses. Hypermutation of the provirus was initially identified by manually inspecting the reads in IGV, the consensus sequence from the regions of the provers with coverage >4 was concatenated, aligned using CLUSTAL O (1.2.4) and checked for hypermutation with Hypermut (https://www.hiv.lanl.gov/content/sequence/HYPERMUT/hypermut.html).

**Endogenous retroviruses**

The sequence of bovine *APOB* ERV was generated by PCR amplifying the full length ERV with LongAmp Taq DNA Polymerase (New England Biolabs) from a Holstein suffering from cholesterol deficiency. The resultant PCR product was sequenced on the Illumina platform as described below. It was also sequenced with an Oxford Nanopore MinION R7 flow cell as previously described[26]. Full length sequence of the element was generated via manual curation. Guide RNAs and primer pairs were designed using this ERV reference. For the Ovine ERV we used the previously published enJSRV-7 sequence[30] as a reference to design PCIP guide RNAs and PCR primers.

As the ovine and bovine genome contains sequences matching the ERV, mapping ERV PCIP-seq reads back to the reference genome creates a large pileup of reads in these regions. To avoid this, prior to mapping to the reference we first used BLAST[45] to identify the regions in the reference genome containing sequences matching the ERV, we then used BEDtools[46] to mask those regions. The appropriate ERV reference was then added as an additional chromosome in the reference.

**PCR validation and Illumina sequencing**

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Clone specific PCR products were generated by placing primers in the flanking DNA as well as inside the provirus. LongAmp Taq DNA Polymerase (New England Biolabs) was used for amplification following the manufacturers guidelines. Resultant PCR products were sheared to ~400bp using the Bioruptor Pico (Diagenode) and Nextera XT indexes added as previously described[26]. Illumina PCIP-seq libraries were generated in the same manner. Sequencing was carried out on either an Illumina MiSeq or NextSeq 500. Clone specific PCR products sequenced on Nanopore were indexed by PCR, multiplexed and libraries prepared using the Ligation Sequencing Kit 1D (SQK-LSK108) and sequenced on a MinION R9.4 flow cell.

**BLV references**

The sequence of the pBLV344 provirus was generated via a combination of Sanger and Illumina based sequencing with manual curation of the sequence to produce a full length proviral sequence. The consensus BLV sequences for the bovine samples 1439 & 1053 were generated by first mapping the PCIP-seq Nanopore reads to the pBLV344 provirus. We then used Nanopolish[47] to create an improved consensus. PCIP-seq libraries sequenced on the Illumina and Nanopore platform were mapped to this improved consensus visualized in IGV and manually corrected.

**Genome references used**

Sheep=OAR3.1

Cattle=UMD3.1

Human=hg19

Sequences of the exogenous and endogenous proviruses can be found in Supplementary File 1

**References**

41. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34,** 3094–3100 (2018).
42. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
43. Killick, R., Fearnhead, P. & Eckley, I. A. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* **107,** 1590–1598 (2012).
44. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14,** 178–192 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

45. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
46. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
47. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nature Methods* **12,** 733–735 (2015).

## Acknowledgements

This work was supported by les Amis de l'Institut Bordet, the Fonds de la Recherche Scientifique (FRS), the International Brachet Stiftung (IBS) and a Télévie Grant to V.H. M.A. holds a Post-doctoral Researcher fellowship of the FRS. K.D. is a Scientific Research Worker of Télévie. Dr. Carine Van Lint kindly provided DNA from the U1 HIV-1 cell line. Computational resources were provided by GIGA and the Consortium des Équipements de Calcul Intensif (CÉCI). We thank Wouter Coppieters, Latifa Karim, Manon Deckers and the GIGA Genomics Platform for sequencing services.

## Author contributions

K.D. conceived and designed the study, K.D and M.A. optimized the method, generated and analyzed data. V.H. developed the bioinformatics pipeline and analyzed data. F.A. contributed to data generation. A.M. and O.H. provided patient materials, P.G., N.A. and F.M. collected and provided animal samples. A.B. and D.B. contributed to data analysis and review of the manuscript. K.D. wrote the first draft, M.A. V.H. A.V., P.G., M.G and C.C. contributed to the final manuscript. A.V. and M.G. supervised the study.

## Conflict of interest

The authors declare no conflict of interest

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

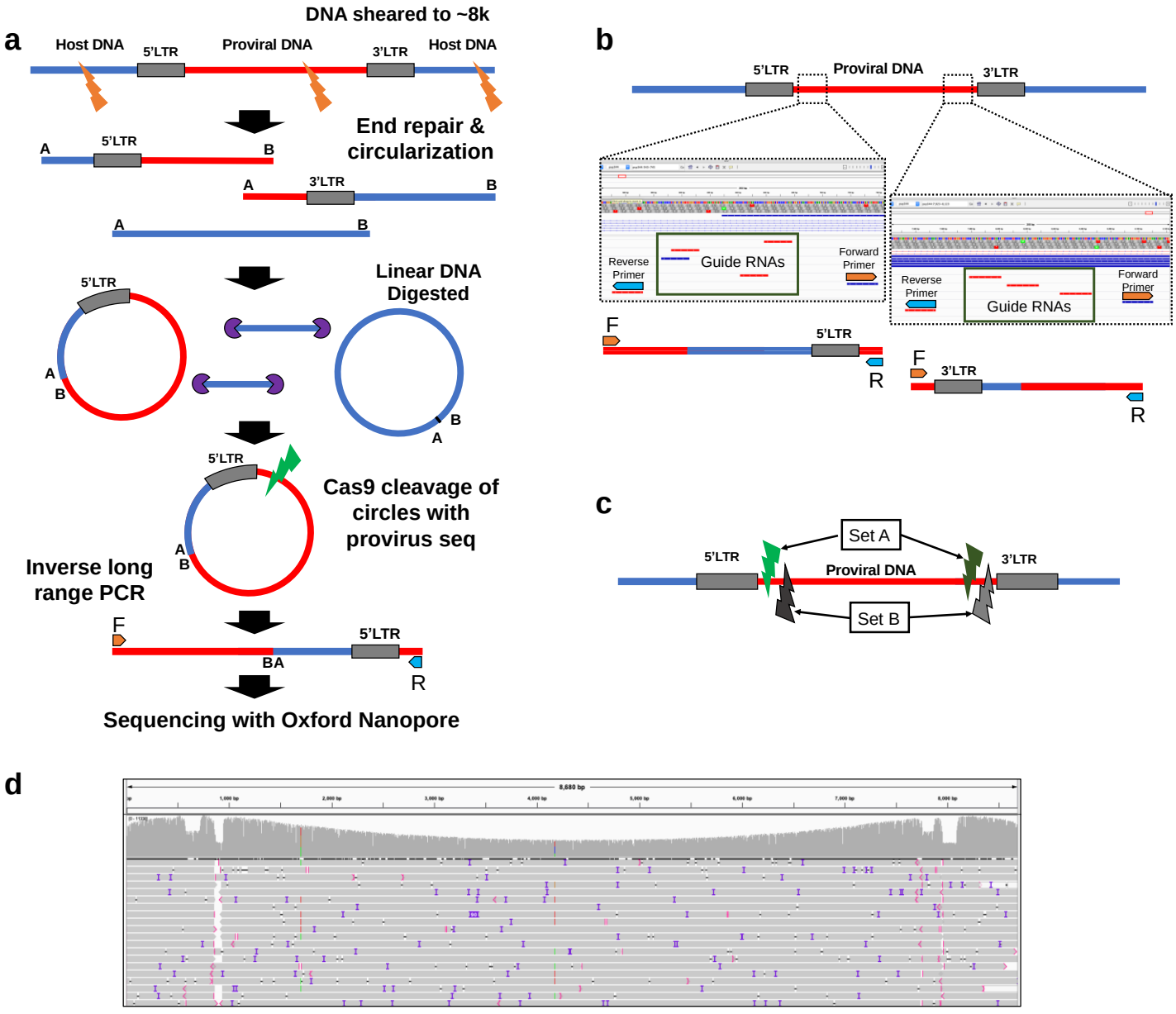


**Figure 1.** Overview of the PCIP-seq method **(a)** Simplified outline of method **(b)** A pool of CRISPR guide-RNAs targets each region, the region is flanked by PCR primers. Guides and primers adjacent to 5' & 3' LTRs are multiplexed. **(c)** As the region between the PCR primers is not sequenced we created two sets of guides and primers. Following circularization, the sample is split, with CRISPR mediated cleavage and PCR occurring separately for each set. After PCR the products of the two sets of guides and primers are combined for sequencing. **(d)** Screen shot from the Integrative Genomics Viewer (IGV) showing a small fraction of the resultant reads (grey bars) mapped to the provirus, coverage is shown on top, coverage drops close to the 5' and 3' ends are regions flanked by primers.

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

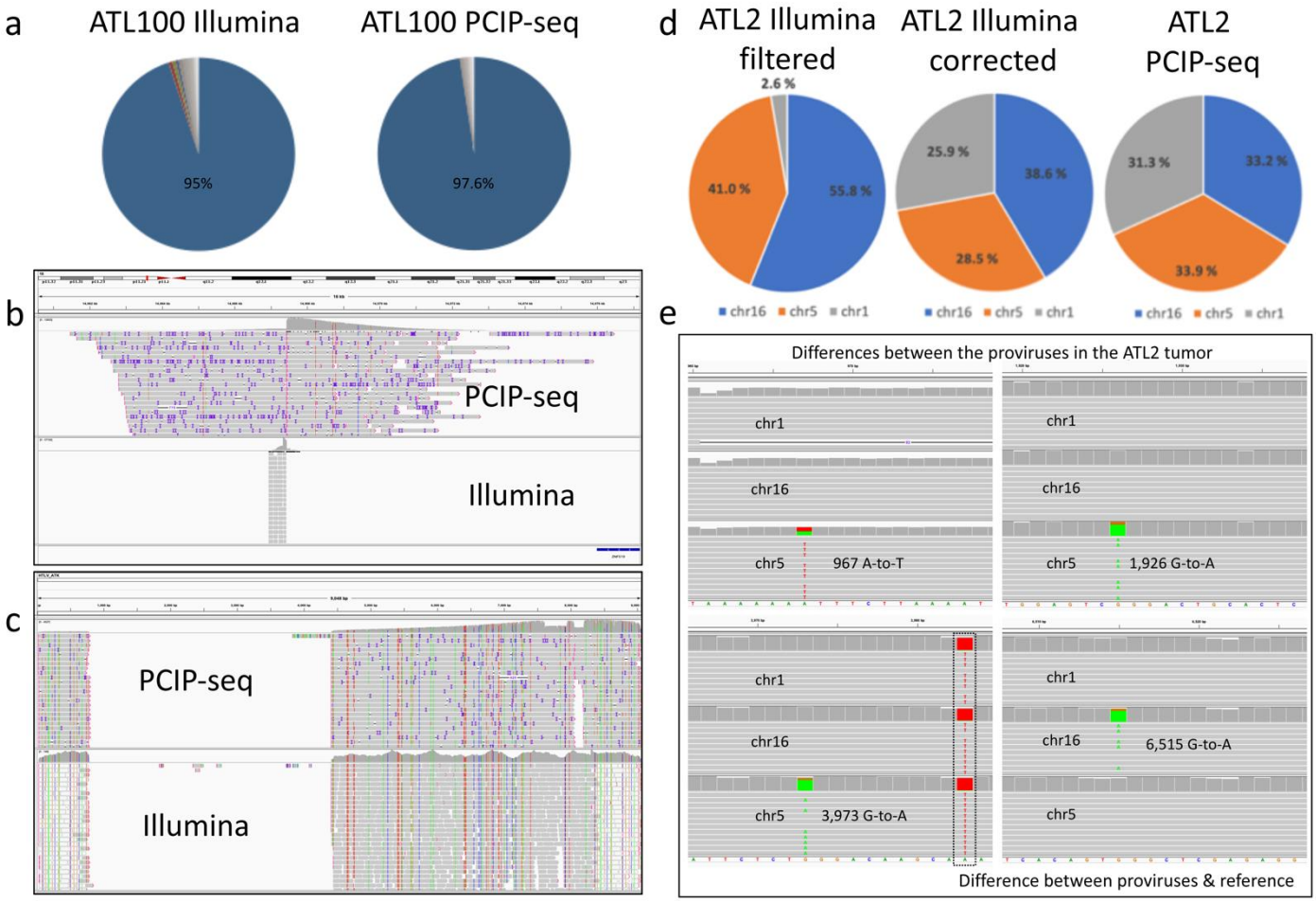


**Figure 2**. PCIP-seq applied to ATL **(a)** In ATL100 both Illumina and Nanopore based methods show a single predominant insertion site **(b)** Screen shot from IGV shows a ~16kb window with the provirus insertion site in the tumor clone identified via PCIP-seq and ligation mediated PCR with Illumina sequencing **(c)** PCIP-seq reads in IGV show a ~3,600bp deletion in the provirus, confirmed via long range PCR and Illumina sequencing. **(d)** The ATL2 tumor clone contains three proviruses (named according to chromosome inserted into), the provirus on chr1 inserted into a repetitive element (LTR) and short reads generated from host DNA flanking the insertion site map to multiple positions in the genome. Filtering out multi-mapping reads causes an underestimation of the abundance of this insertion site (2.6 %), this can be partially corrected by retaining multi-mapping reads at this position (25.9 %). The long PCIP-seq reads can span repetitive elements and produce even coverage for each provirus without correction. **(e)** Screen shot from IGV shows representative reads coming from the three proviruses at positions where four de novo mutations were observed.

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

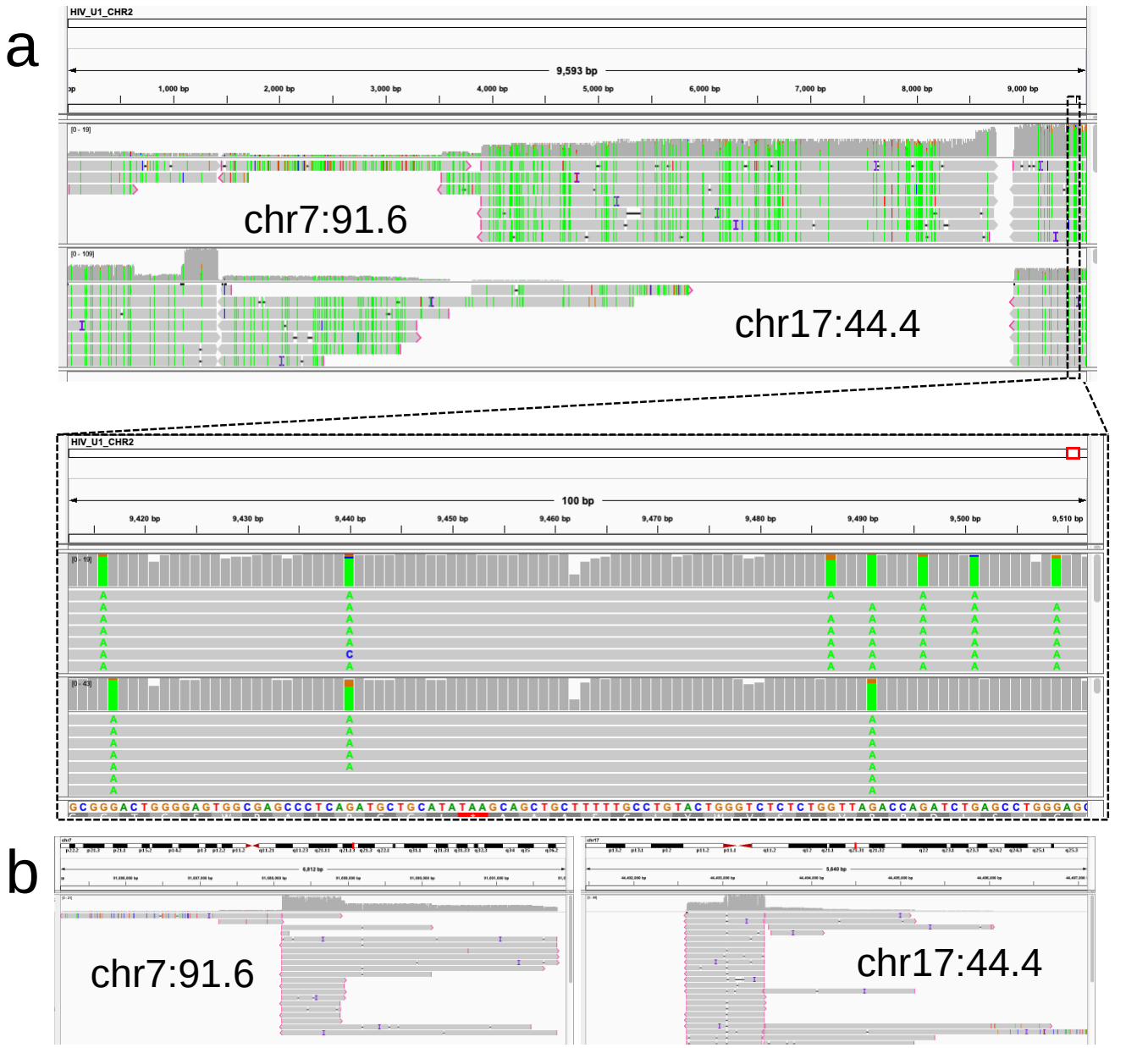


**Figure 3.** APOBEC3G hypermutation in the HIV-1 cell line U1 **(a)** Screen shot from IGV shows reads from the proviruses chr17:44.4 and chr7:91.6 aligned to the sequence of the major U1 chr2 provirus. The vast majority of the de novo mutations observed in both proviruses are G-to-A mutations, the extremely high frequency indicates the action of APOBEC3G. **(b)** The portion of the reads mapping to the host genome shows no evidence of elevated G-to-A mutations, excluding technical artifacts as a source of the variants.

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

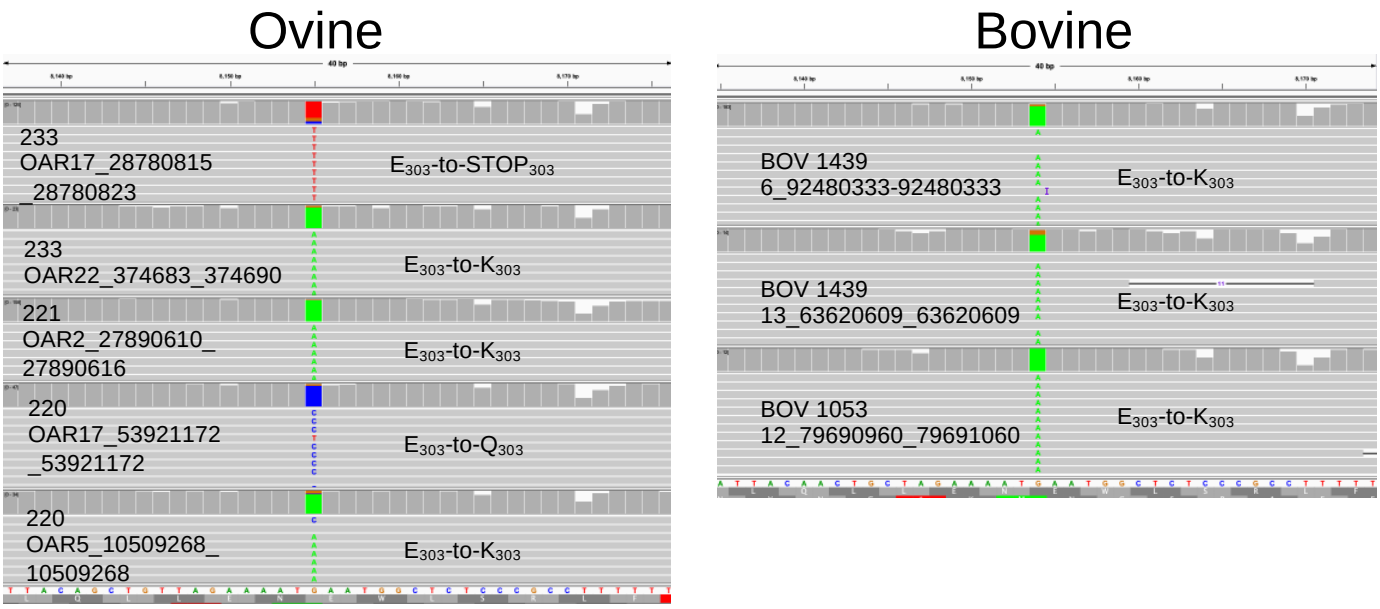


**Figure 4.** Screen shot from IGV shows representative reads from a subset of the clones from each BLV-infected animal with a mutation in the first base of codon 303 in the viral protein Tax.

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

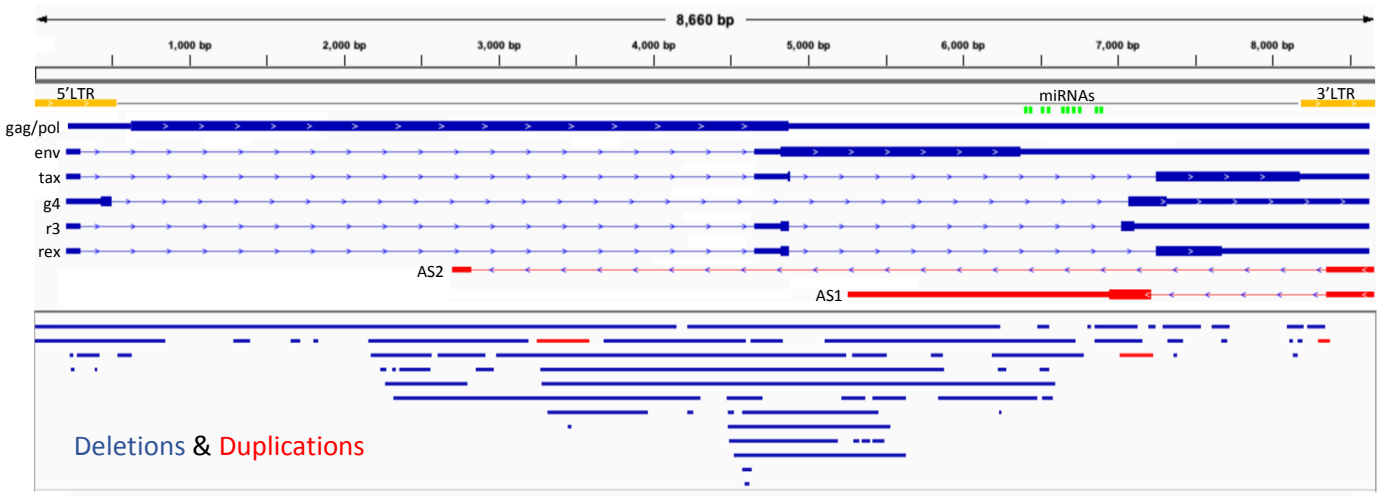


**Figure 5.** Structural variants observed in the BLV provirus. BLV sense and antisense transcripts are shown on top. Deletions (blue bars) and duplications (red bars) observed in the BLV provirus from both ovine and bovine samples are shown below.

bioRxiv preprint doi: https://doi.org/10.1101/558130; this version posted February 24, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

| Sample name | Virus | Host | PVL | Template µg | raw reads | % chimeric reads | Insertion sites | Largest clone (%) |
|---|---|---|---|---|---|---|---|---|
| ATL2 | HTLV-1 | HSA | nd | 4 | 430,060 | 63 | 237 | 34 |
| ATL100 | HTLV-1 | HSA | 106 | 4 | 35,833 | 34 | 235 | 98 |
| HIV_U1 | HIV-1 | HSA | nd | 2 | 506,639 | 50 | 1,172 | 54 |
| 233 | BLV | OAR | 78.3 | 7 | 1,057,945 | 46 | 8,182 | 8.4 |
| 221 (022016) | BLV | OAR | 63.0 | 4 | 391,139 | 59 | 11,379 | 0.7 |
| 221 (032014) | BLV | OAR | 16.0 | 4 | 116,726 | 55 | 8,335 | 0.3 |
| 220 | BLV | OAR | 3.8 | 2 | 99,054 | 58 | 1,921 | 6.4 |
| 1439 | BLV | BosT | 45.0 | 3 | 454,406 | 66 | 7,499 | 1.2 |
| 560 | BLV | BosT | 0.64 | 1 | 92,702 | 41 | 233 | 5.3 |
| 1053 | BLV | BosT | 23.5 | 6 | 627,074 | 62 | 21,579 | 0.4 |

**Table 1** Number of insertion sites (IS) identified via PCIP-seq. Chimeric reads = reads containing host and viral DNA. Largest clone % = insertion site with highest number of reads in that sample. PVL = Proviral Load.