



RESEARCH

Open Access



# Reconstruction of a replication-competent ancestral murine endogenous retrovirus-L

Daniel Blanco-Melo<sup>1,3</sup> , Robert J. Gifford<sup>2</sup> and Paul D. Bieniasz<sup>1\*</sup> 

## Abstract

**Background:** About 10% of the mouse genome is composed of endogenous retroviruses (ERVs) that represent a molecular fossil record of past retroviral infections. One such retrovirus, murine ERV-L (MuERV-L) is an *env*-deficient ERV that has undergone episodic proliferation, with the most recent amplification occurring ~2 million years ago. MuERV-L related sequences have been co-opted by mice for antiretroviral defense, and possibly as promoters for some genes that regulate totipotency in early mouse embryos. However, MuERV-L sequences present in modern mouse genomes have not been observed to replicate.

**Results:** Here, we describe the reconstruction of an ancestral MuERV-L (ancML) sequence through paleovirological analyses of MuERV-L elements in the modern mouse genome. The resulting MuERV-L (ancML) sequence was synthesized and a reporter gene embedded. The reconstructed MuERV-L (ancML) could replicate in a manner that is dependent on reverse transcription and generated *de novo* integrants. Notably, MuERV-L (ancML) exhibited a narrow host range. Interferon- $\alpha$  could reduce MuERV-L (ancML) replication, suggesting the existence of interferon-inducible genes that could inhibit MuERV-L replication. While mouse APOBEC3 was able to restrict the replication of MuERV-L (ancML), inspection of endogenous MuERV-L sequences suggested that the impact of APOBEC3 mediated hypermutation on MuERV-L has been minimal.

**Conclusion:** The reconstruction of an ancestral MuERV-L sequence highlights the potential for the retroviral fossil record to illuminate ancient events and enable studies of the impact of retroviral elements on animal evolution.

## Background

Uniquely among animal viruses, retroviruses integrate into the genome of the host cell as an obligate step in their replication cycle. Because the target cells of some retroviruses can include cells of the germ line, proviruses can occasionally become vertically inherited [1]. A subset of these inherited proviruses can become fixed in the population through genetic drift, or sometimes by providing an evolutionary advantage to the host. Inherited proviruses are termed endogenous retroviruses (ERVs) and are present in all animal species that have been examined, accounting for approximately 8 and 10% of the

human and mouse genomes, respectively [2]. In nearly every case, however, fixed proviruses have inactivating mutations that prevent their further spread.

The vast array of ERVs represent an extensive viral fossil record that provides an opportunity to study the biology of ancient or extinct retroviruses, and the effects that these viruses have had on the evolution of their hosts [3]. Previously, we and others have reconstructed a full-length infectious human ERV (HERV-K) [4, 5], functional capsid proteins of endogenous chimpanzee gammaretroviruses (CERV 1 and 2) [6, 7], and lentiviruses, PSIV and RELIK [8], as well as functional envelope proteins from CERV2 and HERV-T [9, 10]. These reconstruction experiments have enabled the identification of ancient virus receptors [9, 10] and demonstrated the ancient origin of cyclophilin A-lentiviral capsid interactions [8].

\*Correspondence: [pbieniasz@rockefeller.edu](mailto:pbieniasz@rockefeller.edu)

<sup>1</sup> Laboratory of Retrovirology and Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA

Full list of author information is available at the end of the article



Additionally, these studies have shown that the replication of HERV-K, CERV1 and CERV2 was affected by the APOBEC3 cytidine deaminases [5, 6, 11]. Overall these “paleovirological” studies have provided previously inaccessible insights into the co-evolution of viruses and hosts.

Murine ERV-L (MuERV-L) is an abundant mouse ERV that is transcriptionally active at an early (2-cell) stage of the mouse embryo [12–15]. Previous analyses have established that MuERV-L underwent two amplification bursts, one after the divergence of the *Mus* and *Rattus* genera around ~10 million years ago (MYA), and a more recent and prolific burst about 2 MYA, which is distinguished by the presence of a 33nt in-frame deletion in the 5' half of the *gag* ORF (Fig. 1a) [16]. These amplifications led to the deposition of thousands of MuERV-L derived sequences in the mouse genome. Moreover, MuERV-L belongs to a larger family of ERV-L elements that have been active throughout the evolution of mammals [17–19]. In contrast to their human counterparts, many MuERV-L elements have complete coding potential, encoding open reading frames (ORFs) for *gag* and *pol* (Fig. 1a) [20]. However, like other ERV-L elements, MuERV-L is characterized by the complete absence of an *env* gene that, coupled with its highly restricted early transcription profile, suggests an entirely intracellular retrotransposon-like replication cycle [21]. Indeed, MuERV-L transcripts are able to give rise to intracellular viral-like particles that accumulate in the endoplasmic reticulum [13] but are not thought to be replication competent.

MuERV-L related or derived sequences appear to have been co-opted for two distinct biological activities in the mouse. The antiretroviral restriction factor Fv1, which inhibits infection by MuLV and certain other retroviruses is derived from MuERV-L-like Gag sequences and appeared in the mouse genome at least 5 MYA [22–24]. Additionally, recent studies have suggested that the propensity of MuERV-L to be transcriptionally active only at the two-cell stage of mouse embryogenesis may have led to the co-option of its long terminal repeats (LTRs), as promoters of genes involved in the zygotic genome activation [14, 15]. Transcriptional activity of MuERV-L LTRs in two-cell mouse embryos may drive the expression of hundreds of genes that contribute to the totipotency of the blastomeres, but also results in the expression of MuERV-L Gag–Pol polyprotein and the formation of intracellular viral-like particles [13–15]. As development progresses, MuERV-L LTRs appear to be silenced [14, 25, 26]. The expression of MuERV-L at the two-cell stage does not induce an increase in their copy numbers, suggesting that the expressed proviruses do not have the potential to re-integrate into the genome [25].

In fact, no extant copy of MuERV-L has been demonstrated to be capable of completing a replication cycle. Therefore, we set out to derive a replication-competent MuERV-L, based on the premise that an ancestral reconstruction would deliver a sequence that most closely resembles that of a functional ancestor. Herein, we describe the analysis of MuERV-L elements in the mouse genome, a successful reconstruction of a ~2MY old replication-competent ancestral sequence and an analysis of its replication and its interaction with the components of the host intrinsic/innate immune systems.

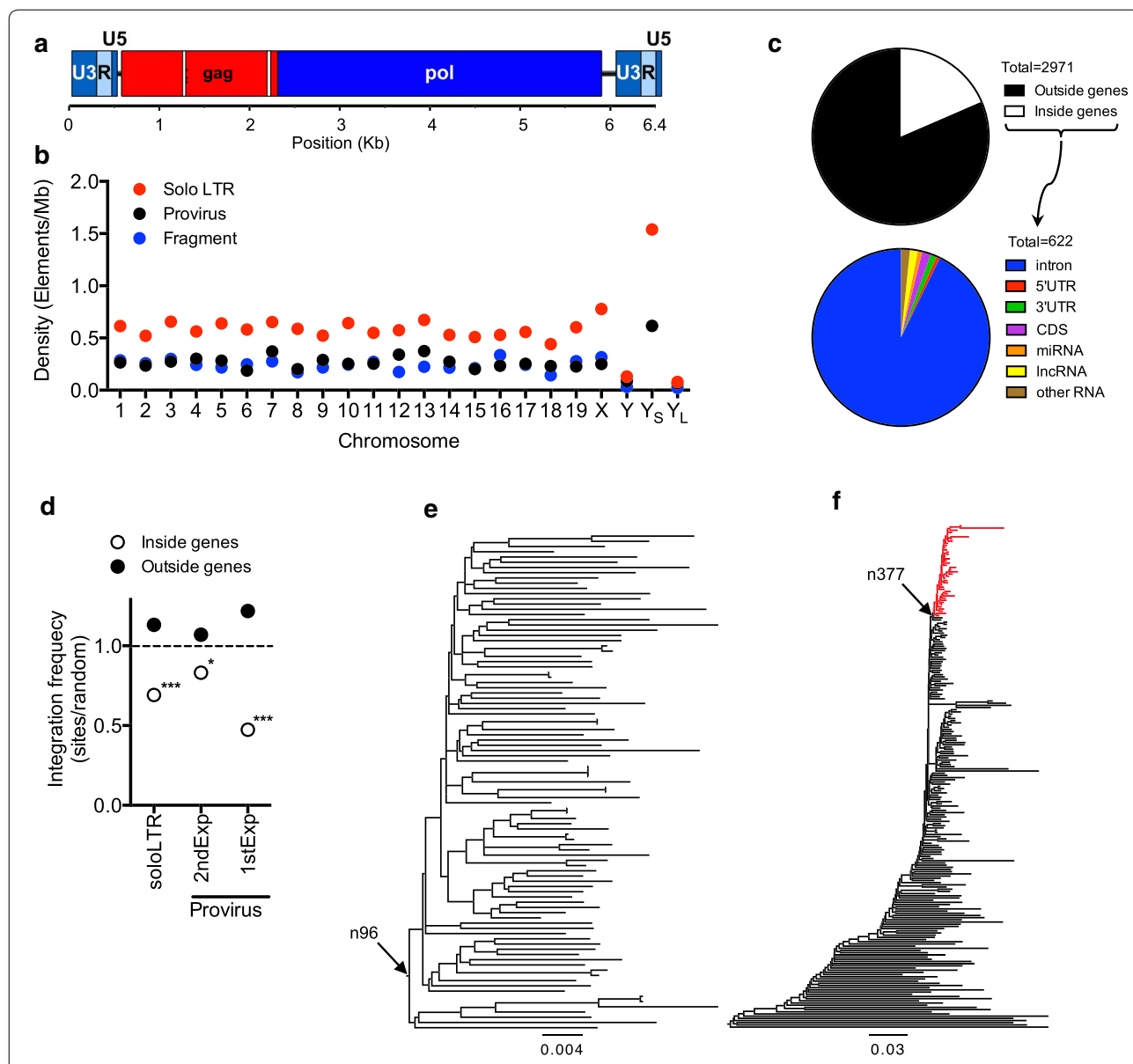
## Methods

### Bioinformatic analyses and ancestral reconstruction

Screening for MuERV-L elements was performed using amino acid and nucleotide sequences from the MuERV-L reference sequence (MuERV-L<sup>ref</sup>, GenBank: Y12713) [20] as probes for tBLASTn (*gag* and *pol*) and BLASTn (LTR) [27] searches of two mouse genome assemblies: Mm\_Celera (NCBI: GCF\_000002165.2) [28] and GRCm38/mm10 (UCSC: mm10) [29]. To avoid the identification of sequences from related but distinct retroviruses, BLAST hits with an e-value  $\leq 1e-10$  were used as probes for a second round of BLASTx or BLASTn searches against a previously constructed database of endogenous and exogenous class III retroviral sequences. Results from these BLAST searches were imported into a relational database to facilitate the management of the screening process and analysis of hits. Reciprocal hits to MuERV-L<sup>ref</sup> were first ordered by chromosome and orientation and then adjacent or overlapping hits were assembled into proviral loci by comparison with the MuERV-L<sup>ref</sup> sequence, allowing for insertions no longer than 10,000 nucleotides. The resulting MuERV-L loci were annotated as genomic features of GRCm38 (downloaded from Ensembl [30] using BioMart) by comparison of their chromosome location using in-house Perl scripts.

Dates of integration for MuERV-L elements were estimated by determining the divergence (K) to a consensus sequence (for solo LTRs) or between paired LTRs (for provirus-containing loci) using PAUP\* [31], divided by  $2 \times$  the mouse neutral substitution rate (r) of  $4.5 \times 10^{-9}$  substitutions per site per year [29] (K/2r) [32, 33]. The mean sequence identity between the consensus LTR sequence and each of the solo LTRs was 95.43%, with very few outliers (1.5% of solo LTRs showed less than 80% identity), suggesting that the consensus sequence is adequate to perform these estimations. Nevertheless, the estimated ages derived using this method represent approximations to the integration dates and should be treated as such.

Statistical analyses of the positions of MuERV-L elements relative to mouse genomic features were



**Fig. 1** Structure and distribution of MuERV-L elements in the mouse genome and reconstruction of an ancestral ~2MY old MuERV-L sequence. **a** Schematic representation of the structure of MuERV-L elements. White boxes in *gag* represent the 33 and 39nt deletions in *gag* at nucleotide positions 671 and 1597, respectively. **b** Distribution of distinct MuERV-L structures among mouse chromosomes.  $Y_5$  = Y chromosome short arm,  $Y_L$  = Y chromosome long arm. **c** Distribution of MuERV-L elements in mouse genomic features (GRCm38/mm10). The fraction of the 2971 elements inside and outside genes is depicted (top chart), as the fraction of elements present in each type of gene or gene feature (bottom chart). **d** Distribution of MuERV-L elements in genic or intergenic regions relative to random controls. The measured value indicates the percentage of MuERV-L elements in each population divided by that of the random controls (see “Methods”). The horizontal dotted line indicates no difference between the ratio of MuERV-L elements in each population and that of the controls. (\*) *p* value < 0.05. (\*\*\*) *p* value < 0.001. *P* values are based on Chi-squared goodness-of-fit or contingency table tests. **e** Maximum likelihood phylogenetic tree of 95 LTR-*gag-pol*-LTR MuERV-L elements in the mouse genome. Arrow denotes the ancestral node reconstructed by baseml (*pol* and LTR, node 96). **f** Maximum likelihood phylogenetic tree of 230 *gag-pol* containing MuERV-L elements in the mouse genome. The monophyletic red clad contains only elements with a 33nt deletion in *gag* at position 671 with or without an additional 39nt deletion in *gag* at position 1597. Arrow denotes the ancestral node reconstructed by baseml (*gag*, node 377)

performed using the Pearson’s Chi-squared test for count data (*chisq.test*) implemented in R. As controls, two sets of random coordinates in the mouse genome were

computationally generated using in-house Perl scripts and the *runif* function implemented in R. For solo LTR comparisons, 10,000 random mouse sequences of 500

nucleotides in length were generated. For proviral loci comparisons, 5000 random mouse sequences of 6500 nucleotides in length were generated. The coordinates of both MuERV-L integrations and random sequences were then mapped to the GRCm38 genome assembly to determine overlapping genomic features (intergenic regions, genes, common repeats and others) using in-house Perl scripts. For ancML integration comparisons, in-house Perl scripts were used to generate controls, consisting of 1000 randomly selected EcoRV-containing fragments (of 1000 nucleotides in length) from the Chinese hamster genome. Each ancML integration site was matched with three of these genomic sequences such that control sites were equidistant from an EcoRV site as was each ancML integration site, as described in [34]. These sites were then mapped to the Chinese hamster genome (criGri1) [35] to determine the overlapping genomic features.

Ancestral reconstruction of MuERV-L was performed using two distinct sequence sets. For the reconstruction of the ancestral *pol* gene and the LTRs, we used a set of 95 complete proviral sequences (LTR-*gag-pol*-LTR) identified by default-parameter based BLASTn searches of GRCm38 using MuERV-L<sup>ref</sup>. Each proviral sequence was individually aligned to MuERV-L<sup>ref</sup> using MUSCLE [36] and a multiple sequence alignment (MSA) was generated using the profile alignment function of MUSCLE. Insertions relative to MuERV-L<sup>ref</sup> were eliminated from the MSA, except a 6nt insertion at position 298 and 6249 in both LTRs that was shared between 25% of the sequences. The MSA was used to construct a maximum likelihood (ML) phylogenetic tree using raxML [37] with the following parameters: rapid bootstrap analysis with 1000 replicates under GTRCAT followed by a ML search under GTRGAMMA to evaluate the final tree topology (-m GTRCAT -# 1000 -x 13 -k -f a). Thereafter, the tree was midpoint rooted. The MSA together with the phylogenetic tree were used to guide an ML ancestral reconstruction using baseml from the PAML package [38] (model: REV, initial values of alpha and kappa were calculated on the MSA by jmodeltest [39], branch lengths were used as initial values). For the ancestral reconstruction of the *gag* ORF we first aligned and constructed a phylogenetic tree using 230 *gag-pol* containing sequences (identified by our screening of the mouse genome described above). We determined the presence or absence of the 33 and 39nt deletions in the *gag* ORF relative to MuERV-L<sup>ref</sup> (that does not show any deletion) and identified a monophyletic clade of 40 sequences that had the 33nt deletion in *gag* (irrespective of the status of the 39nt deletion). Thereafter, reconstruction of the ancestral *gag* sequence corresponding to the internal node for this monophyletic clade was performed as described above. A correction

for the effect of methylation-induced mutations at CpG dinucleotides was applied on both strands of all three sequences (*gag*, *pol* and LTR) as described in [8]. Specifically, if a particular site where the ancestral reconstruction estimated a TG dinucleotide but at least 10% of the sequences in the MSA encoded a CG at that position, the TG state was considered to be the result of methylation-induced mutation and the sequence at this position was assigned as CG. The resulting sequences were combined to produce the sequence from which ancML was derived.

Hypermutation analysis and statistics were performed using Hypermutter 2.0 [40] on the set of 230 *gag-pol* containing sequences used to reconstruct an ancestral *gag*, using either default parameters or with exclusion of sites with a 5' C next to the mutated G.

### Plasmid construction

To construct ancML, sequences from the U3 region of the MuERV-L 5'LTR 5' to the TATA box were substituted with corresponding CMV promoter sequences. We also added an extra 12nt containing two MluI sites immediately 3' to the *pol* stop codon to facilitate the insertion of a reporter gene. The modified ancML sequence was synthesized and inserted into pUC57 (Genewiz, NJ). The replication dependent LINE-1 element (L1.3 plasmid) [41] was kindly provided by Dr. John V. Moran. The replication dependent *neo* cassette (a *neo* gene controlled by a SV40 promoter and interrupted by an intron) was amplified by PCR from the L1.3 plasmid and inserted into ancML using the MluI sites at the 3' end of *pol*. A separate pCR3.1 based plasmid, expressing GFP (from a CMV promoter) and a *neo* gene (NEO, expressed from a SV40 promoter) was used as a control.

The ancML-RTmut construct was created by using overlapping PCR and primers that annealed to the RT active site with four nucleotide mismatches, the PCR fragment was inserted into ancML using unique surrounding BstZ17I and NheI restriction sites contained in the outmost primers, generating an ancML with a mutated RT active site (YIDD to AIAA).

The ancMLΔGAAGT construct was generated using PCR and a reverse primer that annealed to the 5' end of the PBS and the 3' end of the U5 region of the 5'LTR, and lacked the intervening 5nt linker sequence (GAAGT). The PCR fragment was inserted into ancML using unique AgeI and KpnI restriction sites contained in the forward and reverse primers, respectively.

A plasmid expressing mouse APOBEC3 (mA3, C57BL/6J strain) was kindly provided by Rachel Libertore (unpublished). A C-termini HA-tagged version of mA3 was produced by PCR using primers containing two HA tags and a 15nt linker sequence, following previously

published functional human HA-tagged APOBEC3 proteins [42, 43]. This construct was introduced into the retroviral expression plasmid LBCX using unique SfiI sites. A retroviral vector (LBCX) expressing Fv1<sup>bbn</sup> was kindly provided by Dr. Theodora Hatzioannou [44].

#### Cell culture

Cell lines (except CHO-K1 and pgsA cells) were maintained in Dulbecco's Modified Eagle Medium (DMEM), Eagle's Minimum Essential Medium (EMEM) or Roswell Park Memorial Institute medium (RPMI) supplemented with 10% FBS and gentamycin (2 µg/ml, Gibco) according to ATCC instructions. CHO-K1 and pgsA cells were maintained in Ham's F-12 media supplemented with 10% FBS, 1 mM of L-glutamine and 2 µg/ml of gentamycin. All cells were incubated at 37 °C, except DF-1 cells that were incubated at 39 °C.

#### Generation of CHO cell lines expressing murine APOBEC3

293T cells were transfected (using polyethylenimine) with plasmids expressing MuLV gag-pol, and VSV-G, along with an LBCX based retroviral vector expressing HA-tagged mA3 or Fv1<sup>bbn</sup>. Viral stocks were harvested and filtered (0.22 µm) 2 days after transfection, and were used to transduce CHO-K1 cells (seeded in 24 well plates). Transduced cells were expanded in 10 cm dishes with media supplemented with 5 µg/ml of blasticidin (Thermo Fisher Scientific Inc.). Single cell clones expressing mA3 were isolated by seeding blasticidin resistant cells at 0.5 cells per well in a 96 well plate. Three distinct single clones that expressed mA3 in 100% of the cells (tested by immunofluorescence) were used in ancML replication assays.

#### Immunofluorescence assay

Individual clones of CHO cells expressing murine APOBEC3 were fixed with 4% paraformaldehyde (PFA) for 30 min followed by treatment with 10 mM glycine (diluted in PBS) for another 30 min. Cells were permeabilized with a buffer containing 0.1% of Triton X-100 and 5% goat serum (diluted in PBS) for 15 min. Cells were then washed 2 times with PBS before being treated with mouse monoclonal anti-HA antibody (Covance) diluted in a buffer containing 0.1% Tween-20 and 5% goat serum (diluted in PBS) for 2 h at room temperature. Cells were washed three times with PBS before being treated with goat anti-mouse secondary antibody (Alexa Fluor 488 dye, ThermoFisher) diluted in a buffer containing 0.1% Tween-20 and 5% goat serum (diluted in PBS) for 1 h at room temperature. Cells were washed three more times with PBS and fluorescent microscopy images were analyzed using the EVOS FL Cell Imaging System.

#### MuLV infection assay

293T cells were transfected (using polyethylenimine) with plasmids expressing N-tropic or B-tropic MuLV gag-pol, and VSV-G, along with a CNCG based retroviral vector expressing GFP [45]. Viral stocks were harvested 2 days after transfection, filtered (0.22 µm) and were used to infect control or Fv1<sup>bbn</sup>-expressing CHO cells. Two days post infection the percentage of GFP positive population was quantified using the Guava EasyCyte flow cytometer (Millipore).

#### MuERV-L(ancML) replication assays

The cell lines listed in Table 2 were seeded in 12 well plates 1 day before being transfected with 700 ng of plasmids containing L1.3, ancML or a plasmid expressing *gfp* and a *neo* gene, using 4 µl of Lipofectamine 2000 (Thermo Fisher Scientific Inc.) according to manufacturer instructions. Two days after transfection, cells were plated in 6-well plates with G418 selection media (containing concentrations of G418 that were previously calibrated for each cell type). Ten days later, surviving cells were fixed with 4% PFA and colonies were stained using 0.3% crystal violet in 20% ethanol for counting.

Subsequently, the ancML replication assays were routinely done using CHO-K1 cells as follows. CHO-K1 cells were seeded at  $3 \times 10^5$  cells per well in a 12 well plate. One day later the cells were transfected with 1 µg of plasmid DNA, using 3 µl of Transit-CHO supplemented with 0.5 µl of CHO-mojo reagent (Mirus) diluted in Opti-MEM (Gibco). One day later, the cells were expanded on a 10 cm dish with media supplemented with or without AZT (obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH) or mouse IFN $\alpha$  (Pestka Biomedical Laboratories, Inc.). Two days later, cells were plated in a 15 cm dish or three 96 well plates (for analysis of single cell clones) with media supplemented with 1 µg/ml of G418. For controls, 1/1000 of the cells transfected with the NEO plasmid were plated in the 15 cm dish with selection media. Cells in 15 cm dishes were cultured under selection for 10 days before treatment with 4% PFA and colonies were stained with 0.3% crystal violet in 20% ethanol for counting. Single colonies in 96 well plates were monitored and expanded until reaching confluence in a 10 cm dish. Genomic DNA (gDNA) was extracted from  $5 \times 10^6$  cells using QIAamp DNA mini kit (QIAGEN) for analysis of ancML integration (see below).

To determine the fate of the intron interrupting the *neo* gene during ancML replication, gDNA extracted from CHO pools of cells transfected with a plasmid expressing ancML, ancML $\Delta$ GAAGT or an empty vector, was used as template for PCR analysis. Forward and reverse primers were design to anneal to the extreme 5' and 3' ends of the *neo* gene. For all PCRs performed in this study we

used Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific Inc.).

### Integration site analyses

Sites of ancML integration were determined using a universal Genome Walker kit (Clontech). Briefly, gDNA was extracted from expanded single cell clones of CHO cells following transfection with a plasmid containing ancML and selected in G418. The gDNA was digested with EcoRV (New England Biolabs) and ligated to adaptors. Nested PCRs were performed using forward primers that were designed to anneal to the R region of the 3′LTR and the reverse primers to the adaptor sequence, thereby amplifying 3′ flanking sequences. Bands from second round PCR reactions were gel purified and inserted into pCR-Blunt II-TOPO using the Zero Blunt TOPO PCR cloning Kit (Life technologies) for sequencing. To amplify and sequence the 5′ flanking site we use reverse primers specific to the 5′ LTR and designed forward primers that would specifically anneal to the predicted integration site, based on the previously sequenced 3′ flanking sequence. The resulting CHO gDNA sequences were mapped to the CHO genome (criGri1) using BLAT [46] searches on the UCSC genome browser [47]. To account for biases due to location and density of EcoRV restriction sites, we compared the distribution of the 26 ancML integration sites to matched random controls consisting of three random genomic locations that were at the same distance from an EcoRV site as the site found in the flanking CHO DNA sequence for each MuERV-L integration site [34] (see above).

## Results

### Bioinformatic screens for MuERV-L elements in the mouse genome

To construct a replication-competent MuERV-L sequence we first catalogued the diversity of MuERV-L related sequences in the mouse genome. Currently, there are two available complete mouse genome assemblies: the Mouse Genome Reference Consortium build 38 (GRCm38 also known as mm10) corresponding to the C57BL/6J strain [29], and the whole genome shotgun (WGS) assembly (Celera) that corresponds to a mixture of 5 strains (129X1/SvJ, 129S1/SvImJ, DBA/2J, A/J and C57BL/6J) [28]. We mined both genome assemblies using BLASTn and tBLASTn [27] searches with separate *gag*,

*pol* and LTR probes from a MuERV-L reference sequence (GenBank: Y12713) [20]. The resulting hits were defragmented by merging contiguous hits that mapped to the same locus, representing individual elements.

Overall, we found nearly 3000 MuERV-L elements in the mouse genome that had three major types of structures (Table 1, Fig. 1b, Additional file 1: Table S1). One type comprised complete or near complete proviruses, consisting of an internal *gag-pol* region flanked by two LTRs. The frequency with which this structure occurred was highly discrepant in the two genome assemblies, with 220 proviruses being present in the Celera assembly and 719 in GRCm38 (Table 1). It is unclear whether this discrepancy is due to the different mouse strains or the assembly methods used in each genome project. A second type of MuERV-L elements were solo LTRs that typically arise from recombination between the LTRs flanking a provirus, resulting in the complete excision of the internal sequence. This type of element represents the single most abundant ERV type in animal genomes and accounts for >50% of the MuERV-L elements in the mouse genome (Table 1). The third type of MuERV-L structures, representing ~25% of all elements, were composed of internal sequences with or without a single associated LTR (Table 1).

Because the GRCm38 assembly was better supported by external and internal annotations we utilized this data source to analyze the distribution of MuERV-L elements in the genome (Fig. 1b–d). MuERV-L elements were roughly evenly distributed across mouse chromosomes (Fig. 1b), with the exception of the Y chromosome in which MuERV-L was underrepresented. Indeed, there were only a few elements in the Y chromosome. This finding was surprising, given that other ERVs, (e.g. HERV-K) are enriched in the human Y chromosome [48]. However, in contrast to human Y chromosome, the long arm of the mouse Y chromosome is a highly dynamic gene rich region that has frequently expanded and undergone rearrangement over the past ~3 MY [49]. The distribution of MuERV-L elements in this mouse chromosome shows a clear discrepancy between the short and long arms (Fig. 1b). This finding is likely explained by the low recombination rate of the short arm of the mouse Y chromosome, resulting in an enrichment of mobile DNA elements, while recent gene amplification and

**Table 1 MuERV-L sequences identified in mouse genome assemblies**

Assembly	Strain	References	Total	soloLTR	Provirus	Fragments
GRCm38	C57BL/6J	[29]	2971	1588	719	664
Mm_Celera	Mixture	[28]	2768	1775	220	773

rearrangement events in the long arm may have inhibited the fixation of MuERV-L elements.

The majority (79%) of MuERV-L elements were found in intergenic regions (Fig. 1c). About ~19.5% of elements were found inside introns (Fig. 1c), of which the majority (65%) were found in antisense orientation relative to of the corresponding gene, as previously documented for Intracisternal A particles (IAPs) [50]. The remaining elements (1.5%) were found primarily in non-coding RNA genes and untranslated exons (UTRs, Fig. 1c). Only 10 LTRs were found overlapping coding exons. This distribution differs significantly compared to randomized controls ( $p < 0.001$ ). This enrichment of elements outside genes was also apparent if analysis was confined to solo LTRs ( $p$  value  $< 0.001$ , Fig. 1d) or if proviruses from the 2MYA or 10MYA expansions were analyzed separately ( $p$  values  $< 0.05$ , Fig. 1d). Overall, these observations are consistent with an expected selective pressure against retention of MuERV-L elements in genes, which was less evident for younger loci (2nd expansion, Fig. 1d). Interestingly, 10.13% of proviruses implicated in the 10MYA expansion have lost a recognizable ORF (either *gag* or *pol*), in contrast to integrations implicated in the 2MYA expansion where only 5.02% have lost an ORF, consistent with the notion that a modest selection to purge MuERV-L sequences is ongoing.

#### Reconstruction of an ancestral MuERV-L

To reconstruct the ancestral MuERV-L LTRs and the *pol* ORF, we selected 95 complete (LTR-*gag-pol*-LTR) proviruses that were most closely related to a reference MuERV-L sequence (defined by BLAST searches), as this sequence has retained coding potential for both ORFs, has almost identical LTRs, and contained recognizable functional motifs [20]. These sequences were used to guide a maximum likelihood (ML) reconstruction of the root node (*pol* n96) and a pair of identical LTRs (Fig. 1e).

For reconstruction of the *Gag* gene we took a slightly different approach. As the 2nd expansion was the most prolific and we expected that younger integrations would be less divergent from a functional ancestor than older integrations, we selected *gag* sequences that were specific to the most recent expansion (based on the presence of the 33nt in-frame deletion in *gag*). After aligning 230 *gag-pol* containing loci that were present in both Celera and GRCm38 assemblies to the reference sequence (that does not exhibit the deletion), we identified a monophyletic clade of 40 elements that all contained a 33nt in-frame deletion in the 5' half of their *gag* ORFs (Fig. 1f). The *gag* sequences in this clade were selected to guide a ML reconstruction of this internal node (*gag* n377). Thereafter, the combined ancestral LTR, *gag* and *pol* ancestral sequences were corrected for probable errors

derived from deamination of methylated CpG dinucleotides to create a ~2 MY ancestral MuERV-L sequence (ancML, Fig. 2 and Additional file 2) (see “Methods”).

#### ancML is replication-competent and its replication is dependent on a functional reverse transcriptase

To assess the potential replication competence of the reconstructed ancestral MuERV-L sequence, we inserted the full-length ancML sequence into a plasmid vector, replacing the U3 region of the 5' LTR by a CMV promoter to overcome the highly restricted promoter activity of the MuERV-L LTR (Fig. 3a). This design allows for the loss of the CMV promoter sequence after transcription, reverse transcription and integration, resulting in two identical flanking LTRs. A replication-dependent reporter gene, consisting of a *neo* gene controlled by a separate SV40 promoter and interrupted by an intron, was inserted between *pol* and the 3' LTR (Fig. 3a). While the SV40p-*neo* cassette is in reverse orientation relative to ancML transcription, the splice donor and acceptor sites of the intron are in the same orientation as the ancML transcription. Thus, only if ancML undergoes splicing, reverse transcription and integration in the host cell genome the intron is removed and a functional Neo (G418 resistance) protein is expressed. This approach has been used previously to monitor the intracellular replication of retrotransposons [51].

We determined whether transfection of a plasmid harboring ancML could produce G418 resistant colonies in cultured cell lines. For this purpose we tested a set of 13 cell lines: 6 of primate origin, 5 from rodents, 1 from a carnivore and 1 from an avian species (Table 2). Despite efficient transfection in the majority of the cells tested, and the abundant formation of G418-resistant colonies using a control plasmid, the ancML expression plasmid was able to generate G418 resistant colonies only in Chinese hamster ovary K1 (CHO) cells, CHO-derived pgsA cells, and (to a greatly reduced extent) in Vero cells (Table 2 and Fig. 3b). Surprisingly, a control plasmid expressing a LINE1 element containing the same replication dependent *neo* resistant gene (L1.3 plasmid) [41] failed to produce G418 resistant colonies in 4 of the cells tested, including Vero (Table 2). Conversely, all the cell lines tested generated G418 resistant colonies when transfected with a control plasmid expressing an intact *neo* gene.

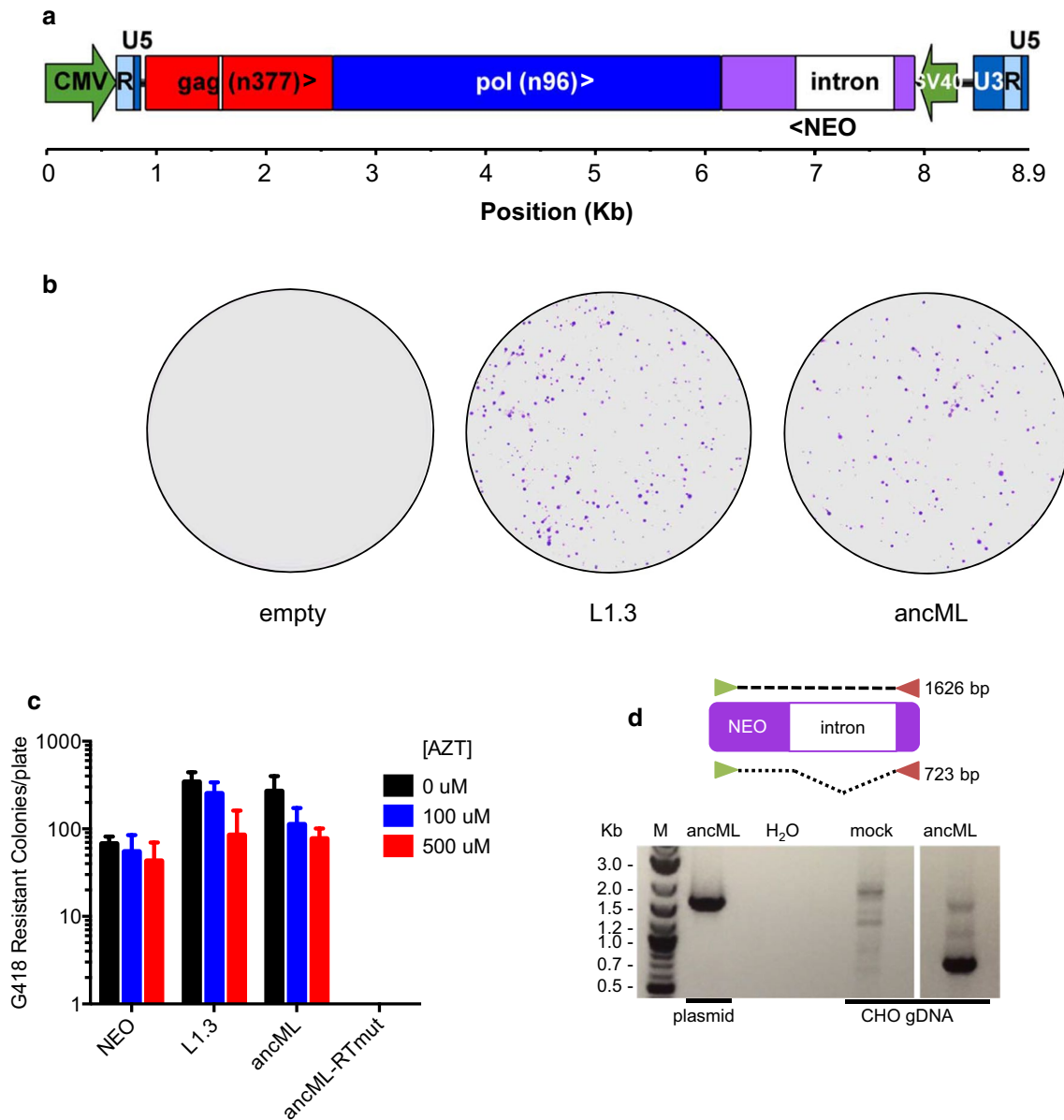
To determine whether the ancML-associated G418 resistant colonies had arisen as a result of ancML replication, we mutated the predicted ancML reverse transcriptase (RT) active site from YIDD to AIAA (Fig. 2). This mutation completely abolished the production of G418 resistant colonies following ancML transfection (Fig. 3c). Additionally, we found that the formation





(See figure on previous page.)

**Fig. 2** Nucleotide sequence and translation products of a reconstructed ancestral MuERV-L LTR sequences are shown in bold italics. Nucleotide and protein sequence of *gag* and *pol* are indicated in red and blue, respectively (amino acid single letter code, (\*) represents stop codons). The 33nt deletion in *gag* is shown with a magenta triangle. The position of the 39nt that are deleted in some MuERV-Ls is highlighted in magenta. The RT active site is highlighted in yellow. The PBS is indicated in violet, the polypurine tract in red, the TATA box in green and the polyadenylation site in bright blue



**Fig. 3** Reverse transcription-dependent ancML replication. **a** Organization of the ancML construct. Green arrows indicate promoter sequences. NEO: *neo* gene in reverse orientation relative to ancML transcription. Chevrons indicate the orientation for each ORF (>: forward, <: reverse). A white box indicates a 33nt deletion in *gag* at position 671. **b** G418 resistant colonies on 15 cm cell culture plates derived from CHO cells transfected with plasmids expressing ancML, L1.3 or an empty vector. **c** Quantification of G418 resistant CHO cell colonies following transfection and treatment in the presence or absence AZT. CHO cells were transfected with plasmids expressing a *neo* gene (NEO), L1.3, ancML or an ancML construct with inactivating mutations in the RT active site (ancML-RTmut). AZT treatment was applied for 2 days before G418 selection. Data are mean  $\pm$  SD from 3 independent experiments. **d** PCR amplification of the *neo* gene in genomic DNA (gDNA) extracted from CHO cells following transfection with a plasmid expressing ancML or an empty vector. A scheme of the PCR amplification strategy is shown on top. The use of template DNA from plasmid or from CHO gDNA as well as a water control is indicated. M: molecular weight ladder

**Table 2 Cell lines tested for replication of ancML**

Cell Line	Organism	Lineage	GFP <sup>#</sup>	NEO <sup>a</sup>	L1.3 <sup>b</sup>	ancML <sup>c</sup>
DF-1	<i>Gallus gallus</i>	Aves	***	+++	++	–
CRFK	<i>Felis catus</i>	Carnivora	**	+++	+	–
CV-1	<i>Cercopithecus aethiops</i>	Primates	*	++	+	–
Vero	<i>Cercopithecus aethiops</i>	Primates	***	+++	–	+
HT1080	<i>Homo sapiens</i>	Primates	**	++	+	–
HOS	<i>Homo sapiens</i>	Primates	*	++	–	–
Huh7.5	<i>Homo sapiens</i>	Primates	*	++	–	–
HeLa	<i>Homo sapiens</i>	Primates	***	+++	++	–
pgsA745	<i>Cricetulus griseus</i>	Rodentia	**	++	+	++
CHO-K1	<i>Cricetulus griseus</i>	Rodentia	**	++	+	++
MusDunni	<i>Mus dunni</i>	Rodentia	***	+++	+	–
SC-1	<i>Mus musculus</i>	Rodentia	**	++	–	–
NIH3T3	<i>Mus musculus</i>	Rodentia	**	++	++	–

All cell lines were transfected using lipofectamine 2000

<sup>#</sup> A plasmid expressing GFP was utilized as a transfection control. Percentage of GFP positive cells: (–) 0%, (\*) < 10%, (\*\*) 10–50%, (\*\*\*) > 50%

<sup>a</sup> A pCR3.1 plasmid expressing a *neo* gene from a SV40 promoter was used as a control for G418 resistant colony formation

<sup>b</sup> The replication dependent LINE1 (L1.3) plasmid [41] was used as a control for retrotransposition and G418 resistance from the interrupted NEO reporter cassette

<sup>c</sup> ancML correspond to a pUC57 plasmid containing the cassette depicted in Fig. 3a

Number of G418 resistant colonies: (–) None, (+) ≤ 10, (++) > 10, (+++) > 50

of G418 resistant colonies by the ancML and L1.3 constructs was modestly reduced in the presence of azidothymidine (AZT), a retroviral RT inhibitor (Fig. 3c), while the formation of G418 resistant colonies by CHO cells transfected with a control plasmid expressing the *neo* gene was nearly unaffected. We isolated genomic DNA (gDNA) from CHO cells that had been transfected with the ancML plasmid and selected in G418, and determined the fate of the intron in DNA forms of ancML by PCR (Fig. 3d). In CHO cells transfected with the ancML plasmid, the vast majority of the amplified DNA sequences corresponded to the properly processed *neo* gene, with the intron excised (Fig. 3d).

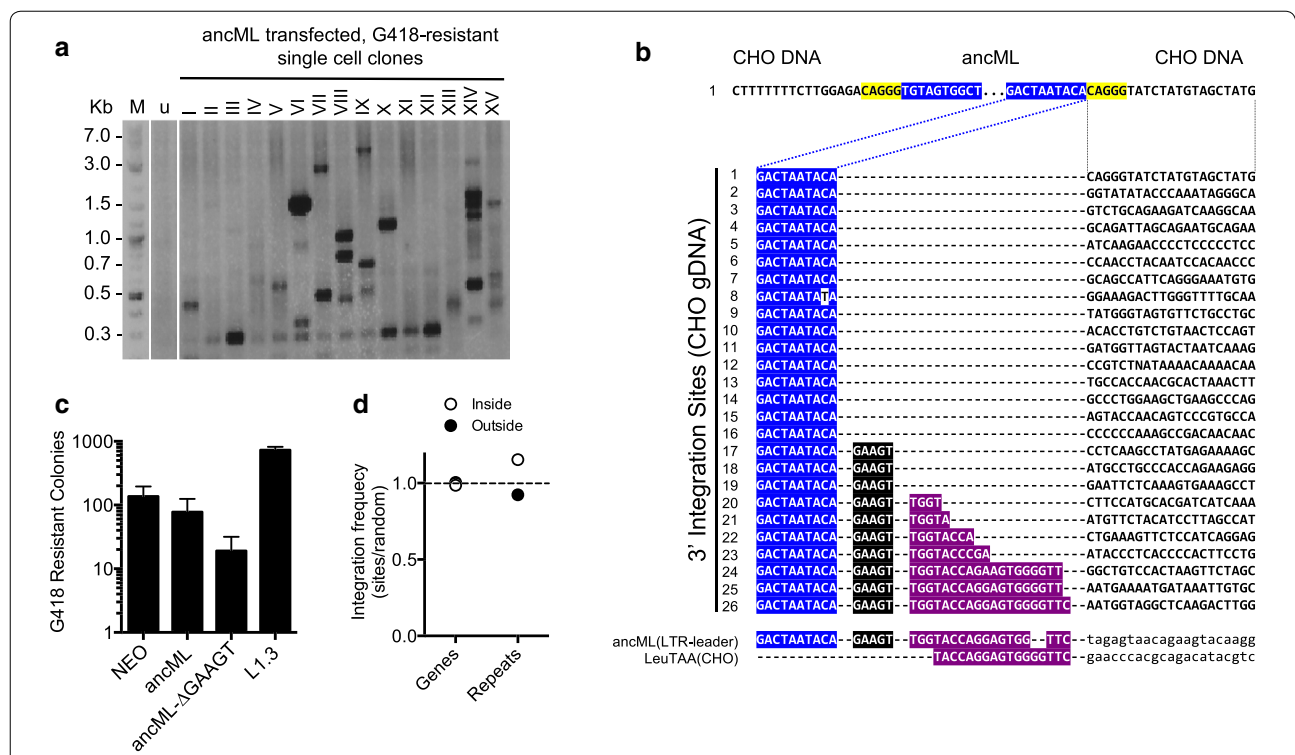
Overall, these results indicate that the reconstructed ancestral MuERV-L sequence is replication competent and able to undergo transcription and reverse transcription upon transfection into CHO cells.

#### Analysis of ancML integration in CHO cells

We next determined whether ancML underwent bona fide integration into CHO cell DNA. For this purpose we used an adapter ligation-PCR technique (Genome Walker kit, Clontech) to amplify integration sites using primers specific to the MuERV-L LTR and an adaptor sequence. The risk of amplifying CHO genomic DNA from hamster ERV-L LTR sequences was minimal as these LTR sequences are quite divergent from those of ancML (only 55.8% of sequence identity) with indels and substitutions in the annealing sites for the primer sets

used. Additionally, hamster ERV-L elements exist only in moderate copy numbers in the Chinese hamster genome [52]. Therefore, DNA from G418-resistant single cell clones of CHO cells, previously transfected with ancML, was digested with restriction enzymes, ligated to linkers and subjected to nested PCR reactions. This procedure revealed bands of varying sizes, sometimes consistent with multiple integrations per cell (Fig. 4a). We cloned and sequenced some of these PCR products. Although some clearly resulted from amplification of the transfected ancML plasmid DNA, we were able to identify 26 *bona-fide* integration sites with CHO genomic DNA flanking the 3' LTR (Fig. 4b and Additional file 3: Table S2). Amplification of sequences flanking the 5' LTR for one of these integrants revealed a five-nucleotide target site duplication (Fig. 4b), as was observed for MuERV-L sequences present in the mouse genome.

Surprisingly, 10 of the 26 3' integration sites included a portion of the 5' leader sequence containing various lengths of the primer-binding site (PBS) and a five-nucleotide LTR-PBS linker sequence, flanking CHO DNA (Fig. 4b). This separation of LTR and PBS is uncommon in exogenous retroviruses and has only previously been observed in another ERV, HERV-E [53]. Elimination of these five nucleotides in the ancML sequence resulted in a ~4 fold reduction in the number of G418 resistant colonies, suggesting an enhancing, but non-essential role for the five nucleotide linker in MuERV-L replication (Fig. 4c). Intriguingly, ~6.5% of the complete



**Fig. 4** Integration of ancML into CHO cell DNA. **a** Example of a genome walker experiment to determine the 3' flanking sequence of ancML integration events in 15 single cell clones that became resistant to G418 following transfection with ancML. Nested PCR reactions were done using EcoRV digested, adapter ligated, gDNA from single G418-resistant cell clones. Forward and reverse primers were designed to anneal to the R region of the 3'LTR and to the adaptor sequence, respectively. M: molecular weight ladder. u: CHO DNA without an integrated ancML insertion. **b** Top: The sequence of an integration site with both 5' and 3' flanking CHO gDNA. The five-nucleotide target site duplication is indicated in yellow. Bottom: Sequences of 26 ancML integration sites in the CHO genome. Sequences of the ancML U5-PBS region as well as the Leucine (TAA) tRNA sequence are included at the bottom of the diagram. Sequence from the U5 region of the 3' ancML LTR is indicated in blue. The 5nt linker sequence is indicated in black. The PBS sequence is indicated in purple. CHO genomic sequences are indicated in bold. Dotted lines indicate correspondence of each sequenced 3' integration junction to the integration site at the top. **c** Enumeration of G418 resistant colonies of CHO cells transfected with plasmids expressing a *neo* gene (NEO), L1.3, ancML and an ancML construct with a deletion of the 5nt linker sequence between the 5' LTR and the PBS (ancML  $\Delta$ GAAGT). Data are mean  $\pm$  SD from 3 independent experiments. **d** Distribution of ancML integration sites in genic, intergenic, or repeat regions relative to matched random controls. The measured value indicates the percentage of ancML integration sites in each population divided by that of the matched random controls (each integration site was matched to three random genomic sequences equidistant to the EcoRV site where the adaptor was ligated). The horizontal dashed line indicates no difference between the frequencies of ancML integration sites in each population compared to the matched controls

(LTR-gag-pol-LTR) proviruses in the mouse genome also contain similar sequences (5-nt linker/PBS) at the end of the 3' LTR, thus showing that this phenomenon also occurred during ancient MuERV-L replication events (Additional file 1: Table S1). During reverse transcription, after the synthesis of the plus-strand strong-stop DNA (+sssDNA), RNase H should remove the primer tRNA, thereby exposing sequences on the +sssDNA that are complementary to the minus strand PBS which will guide the second strand transfer [54, 55]. Inefficient removal of the tRNA primer might result in the synthesis of +sssDNA that includes additional sequences 3' to the PBS. Such a scenario might explain the unusual

integration site structure that we observed for some MuERV-L and ancML insertions (Fig. 4b).

We mapped the position of the 26 ancML integration sites to the Chinese hamster genome using the UCSC genome browser (Fig. 4d) [35, 47]. The Chinese hamster genome (CriGri\_1.0) is currently assembled to the scaffold level and has been annotated by distinct de novo, expression-based and homology gene prediction systems [35]. The majority of the ancML integration sites (19/26 sites) corresponded to intergenic regions, 5/26 sites corresponded to introns and one corresponded to exon 3 of *Znf462*. The single remaining site could not be classified as intergenic or in genes because it mapped to multiple

scaffolds. Of the 26 integration sites, 10 were in elements corresponding to SINE (4), LINE (4) and ERV-L (2) elements. This distribution of ancML integration sites, i.e. within genes versus intergenic regions, as well as within versus outside repetitive sequences, did not differ significantly from matched randomized controls ( $p$  value = 0.97 and 0.56 respectively) (Fig. 4d). Although the distribution of the sequenced ancML integration sites and the distribution of MuERV-L elements in the mouse genome appeared different (Figs. 1d and 4b), our ancML integration site dataset was too small to establish statistical significance. Nonetheless, our results suggest that ancML integration sites are random (or close to random) in their distribution in CHO DNA, in contrast to the distribution of MuERV-L proviruses that are found in the mouse genome which have been subject to selection.

#### ancML is sensitive to innate host antiviral defenses

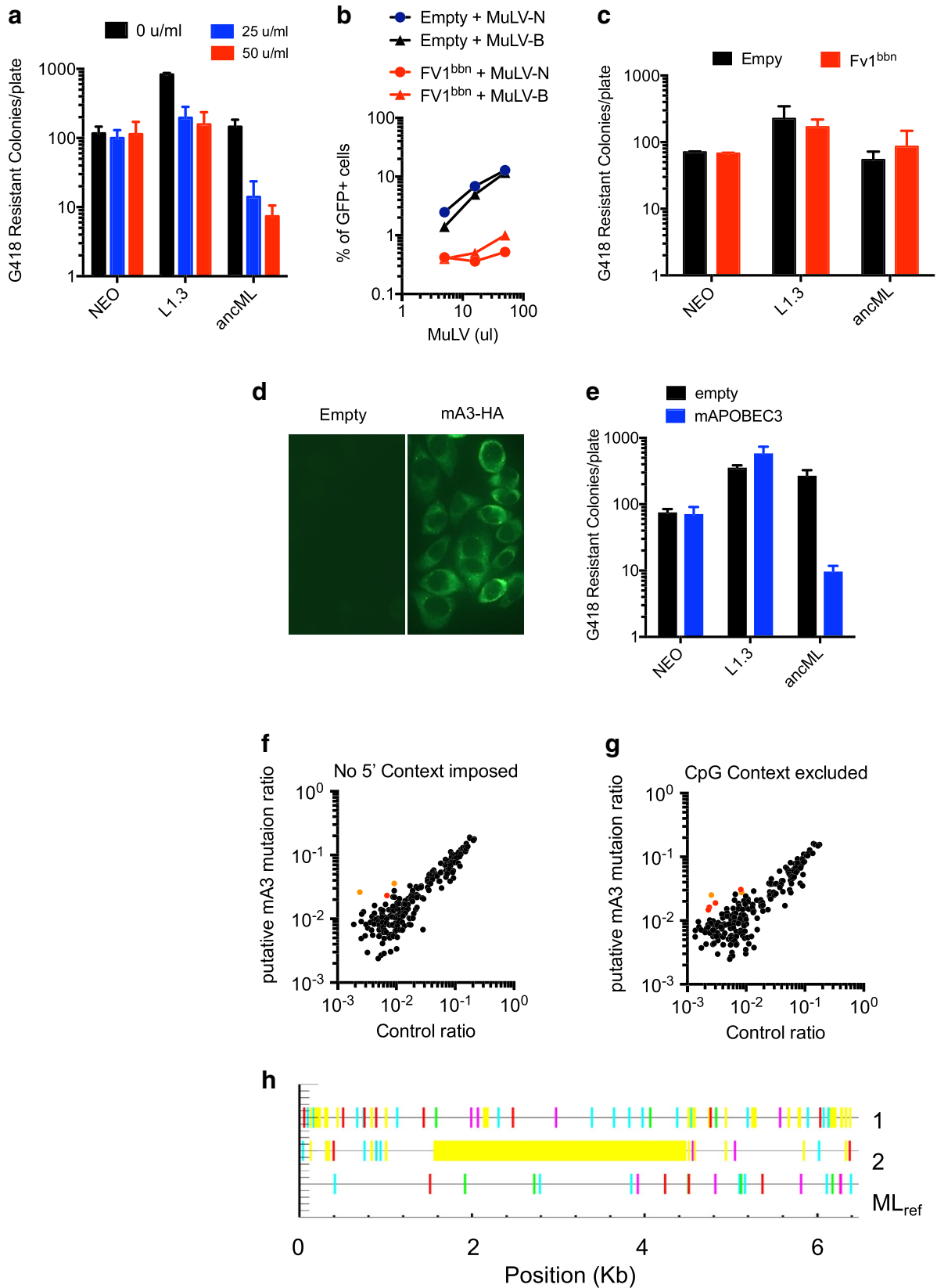
In response to exogenous microbial threats, hosts have evolved sets of genes that sense, and directly interfere with replication of pathogens. One class of such genes which are expressed in response to viral infection following induction by interferons (IFNs), cause a so-called antiviral state [56]. It is not known whether IFNs can inhibit the intracellular replication of retrotransposons. To determine whether ancML replication is affected by type-I IFNs, we transfected CHO cells with plasmids expressing ancML, L1.3 or a *neo* gene and cultured them with media containing varying amounts of murine IFN $\alpha$  (mIFN $\alpha$ ) for 2 days prior to selection in G418 (Fig. 5a). The replication of L1.3 was reduced by ~4 fold upon mIFN- $\alpha$  treatment, and there was a larger, dose dependent effect on ancML, reaching a ~20-fold reduction in

G418 resistant colony formation with 50U/ml of mIFN $\alpha$  (Fig. 5a). Notably, generation of G418-resistant colonies by transfected, non-replicated DNA was not affected by mIFN $\alpha$ . Previous studies have observed that mouse IFN $\alpha$  can stimulate an antiviral state in CHO cells [57–60] and promote the induction of hamster ISGs [61]. Thus, these experiments suggest that IFN $\alpha$  is able to inhibit one or more steps in MuERV-L replication. Interestingly pluripotent stem cells have been shown to express a subset of ISGs [62], and suppression of ERV replication may be one impetus for the acquisition of this property.

We also tested whether specific candidate innate immune effectors could inhibit ancML replication. We first tested if the murine restriction factor Fv1 (thought to have been co-opted from a MuERV-L-like element) could have had an impact on MuERV-L replication. For this we constructed CHO cells stably expressing a chimeric form of Fv1 that shows an expanded resistance to different MuLVs (Fv1<sup>bbn</sup>) [63]. As expected, Fv1<sup>bbn</sup>-expressing CHO cells exhibited resistance to infection by both N-tropic and B-tropic MuLV (Fig. 5b). However, Fv1<sup>bbn</sup>-expressing CHO cells supported ancML, or L1.3 replication (Fig. 5c), at levels similar to those of control cells, indicating that ancML is insensitive to this Fv1 protein. We also tested the ability of mouse APOBEC3, that has been previously shown to inhibit endogenous and exogenous retroviruses (reviewed in [64]), to inhibit ancML replication. For this purpose, we generated CHO cell clones that stably expressing the mouse *Apobec3* in 100% of the cells (Fig. 5d). Remarkably, mouse *Apobec3* (mA3) was able to inhibit ancML replication, reducing G418 resistant colony formation by ~30-fold, but did not affect L1.3 replication (Fig. 5e). The inability of mA3 to

(See figure on next page.)

**Fig. 5** MuERV-L(ancML) replication can be inhibited by innate immune effectors. **a** Enumeration of G418 resistant colonies generated in the presence of increasing amounts of mouse IFN $\alpha$ . CHO cells were transfected with plasmids expressing a *neo* gene (NEO), L1.3, or ancML and cultured with increasing amounts of mouse IFN $\alpha$  for 2 days before G418 selection. Data are mean  $\pm$  SD from 3 independent experiments. **b** Infectivity of MuLV on CHO cells expressing Fv1<sup>bbn</sup>. Percentage of MuLV infected (GFP positive) cells in CHO cells stably expressing Fv1<sup>bbn</sup> (red) or an empty vector (black). Circles and triangles indicate infection by N-tropic or B-tropic MuLV, respectively. **c** Enumeration of G418 resistant colonies of CHO cells expressing Fv1<sup>bbn</sup> or an empty vector were transfected with plasmids expressing a *neo* gene (NEO), L1.3, or ancML. Data are mean  $\pm$  SD from 2 independent experiments. **d** Representative images of Immunofluorescence assays on CHO cells stably expressing an HA-tagged version of mouse APOBEC3 or an empty vector. CHO cells were fixed with 4% PFA and stained with anti-HA antibodies. **e** Enumeration of G418 resistant colonies of CHO cells expressing mouse APOBEC3. Three clones of CHO cells expressing HA-tagged mA3 or an empty vector were transfected with plasmids expressing a *neo* gene (NEO), L1.3, or ancML. Data are mean  $\pm$  SD from 3 experiments with independent single cell clones. **f** and **g** Analysis of MuERV-L elements using Hypermut 2.0. Ratio of G to A mutations at preferred mA3 editing sites (RD 3' to a G) (Y-axis) plotted against ratio of G to A mutations at disfavored mA3 editing sites (YN|RC 3' to a G) (control ratio, X-axis). No 5' context was imposed (**f**), or sites with a 5' C to the mutated G were excluded (**g**). 230 gag-pol containing MuERV-L elements in the mouse genome were compared to their consensus sequence. Data points in red and orange indicate MuERV-L sequences that were statistically significantly enriched in putative mA3 induced mutations ( $p$  value < 0.05). Data points in orange represent MuERV-L elements that are statistically significantly enriched mA3-induced mutations in both analyses ( $p$  value < 0.01). **h** Profile of G to A transitions in two putatively mA3-edited MuERV-L proviral sequences compared to a consensus sequence. The profile of the reference MuERV-L sequence is shown for comparison (ML<sub>ref</sub>, non significantly mA3 edited). Lines in red and cyan represent putative mA3-derived G to A transitions, not accounting for the +2 position (dinucleotide changes from GG to AG and GA to AA respectively), whereas lines in green and magenta represent non mA3-derived G to A transitions (GC to AC and GT to AT respectively). Lines in yellow indicate gaps compared to the consensus sequence



restrict human L1.3 retrotransposition has been previously documented [65], while some human APOBEC3 proteins inhibit L1.3 retrotransposition [42, 66, 67], suggesting that species-dependent differences exist in the ability of APOBEC3 proteins to inhibit the replication of endogenous retroelements.

Because mA3 clearly inhibited ancML replication, and therefore might have affected MuERV-L sequence or replication in vivo, we inspected the 230 gag-pol containing MuERV-L elements that were used to derive ancML gag (Fig. 1f) using Hypermut 2.0 [40] (Fig. 5f–h). For each MuERV-L element we compared the number of G to A transitions in mA3-preferred motifs (5' G(A|G)(A|G|T) 3') with those in control sites (5' G(C|T)N 3' or 5' G(A|G) C 3') relative to a consensus sequence (Additional file 4: Table S3). Only three MuERV-L elements showed significant ( $p < 0.05$ ) evidence of mA3 dependent hypermutation when no 5' context was enforced (Fig. 5f). Because spontaneous deamination of methylated CpG dinucleotides can also produce G to A transitions, we performed the same analysis after excluding sites containing a C nucleotide 5' to the mutated G. When these sites were excluded, 10 MuERV-L elements showed a significant ( $p < 0.05$ ) evidence for mA3 dependent hypermutation (Fig. 5g). Only two MuERV-L elements exhibited statistically significant evidence of mA3 dependent hypermutation in both analyses ( $p$  value  $< 0.01$ ), and both of these elements carried a relatively low mutational burden (Fig. 5f–h). Thus, although ancML replication can be inhibited by mA3, analysis of MuERV-L proviruses in the mouse genome suggests that MuERV-L either rarely encountered mA3, or is inhibited in a manner that prevents the deposition of hypermutated proviruses.

## Discussion

Here, we report the successful reconstruction of a ~2MY old replication competent ancestral MuERV-L sequence, through the analysis of a recently expanded subset of fossilized MuERV-L elements in the mouse genome. According to previous studies [16], and corroborated here, MuERV-L originated ~10 MYA, after the *Rattus–Mus* split and underwent a prolific expansion ~2 MYA. In fact, almost 65% of solo LTRs and MuERV-L proviruses identified herein have an estimated integration date of  $< 3$  MYA. Furthermore, the estimated dates of solo LTRs follow a bimodal distribution (a major one centered ~3MYA and the other ~8MYA) consistent with the estimated times of both expansions (Additional file 1: Table S1). A combination of homology searches and defragmentation methods provided the material for the estimation of the sequence of the ~2MY old replication-competent ancestor.

Other highly abundant *env*-defective ERVs typically appear to be derived from closely related elements that

possess an *env* gene. While other closely related elements do possess an *env* gene, there are no documented ERV-L elements that encode an *env*. It is likely, therefore, that an ancestral ERV-L element lacked an *env* gene. Thus, the bulk of MuERV-L replication likely occurred through entirely intracellular retrotransposon-like mechanisms [21]. Moreover, the bulk of MuERV-L replication likely occurred in early embryos, as the expression of MuERV-L elements appears to be restricted to the 2-cell embryo, although it is unknown whether this property is confined to the subset of elements that proliferated ~2MYA. It is possible that the early embryonic environment is also necessary in some other way for MuERV-L replication given its apparently restricted tropism in cell lines. In particular, it is intriguing that (and as yet unexplained why) MuERV-L only replicated with reasonable efficiency in Chinese hamster ovary cells, even when provided with a promoter that should drive its expression in nearly any cell type.

MuERV-L belongs to an ancient mammalian ERV family (which originated  $> 100$  MYA [18]) that is distantly related to spumaviruses. Therefore, modern functional viral sequences are therefore not useful for attempts to increase the replicative efficiency of ancML. Remarkably, there is a high number of MuERV-L proviruses that have retained their coding potential, and share a high degree of sequence similarity to the functional ancML (with only few coding differences and overall nucleotide identity ranging from 96.16 to 99.31%). However, currently there is no evidence that the ongoing expression of MuERV-L elements at the two-cell stage of the mouse embryo results in successful re-integration, although it is possible that MuERV-L replication and reintegration occurs in modern mouse embryos at some very low rate. Nevertheless, examination of recent *bona fide* integrations might highlight important residues that might be altered to improve ancML replication and/or integration.

We found that mouse IFN $\alpha$  was able to inhibit ancML replication, suggesting that interferon stimulated genes can directly inhibit MuERV-L replication, possibly leading to its recent extinction as a replication competent entity. Alternatively, early embryos may express antiviral proteins that inhibit re-integration of modern MuERV-L elements that would otherwise be intrinsically replication competent [62]. We found that mouse APOBEC3 inhibits ancML replication, but mutational profiles of MuERV-L elements in the mouse genome provide minimal evidence for mA3-dependent hypermutation as a mechanism for inhibition in vivo. During mouse development, mA3 is expressed at the two-cell stage, increasing at the four-cell stage to become one of the top 30% most highly expressed genes [68]. Thus, it is at least possible that mA3 may have acted on replicating MuERV-L elements, perhaps in part through deaminase-independent mechanisms [69].

Despite the apparently random integration pattern of ancML, the analysis of fixed MuERV-L elements showed that there has been a selective pressure to eliminate MuERV-L integrations from genes. Conversely, MuERV-L related sequences (Fv1) have clearly been positively selected to provide defense against retroviral infection [22–24] and recent studies have suggested that regulatory elements of MuERV-L LTRs may have been co-opted to regulate the expression of numerous genes during embryogenesis [14, 15]. While Fv1 arose at least ~5–7 MYA, it is unclear whether the potential exaptation of MuERV-L regulatory sequences occurred during the 10MYA expansion or the more recent ~2 MYA expansion. Nonetheless, there appears to be both a benefit (co-option for antiviral defense and regulation of embryogenesis) and cost (disruption of gene function) associated with the presence of MuERV-L elements in the mouse genome.

MuERV-L appears to be the only member of the ERV-L family that seems to have been reactivated in recent evolutionary times. It is particularly intriguing that the recent expansion is characterized by an in-frame deletion in *gag*, as it could be this deletion the responsible for releasing some MuERV-L elements from the deleterious effects of a hypothetical inhibitory factor ~2MYA. Recent studies have shown the fundamental role that some endogenous retroviral sequences may play in mammalian development and protection from exogenous retroviral infection [15, 23, 24, 70–73]. Indeed one report has suggested that knockdown of MuERV-L transcripts impacts embryonic development [74]. Nevertheless, it remains to be determined whether the current presence of MuERV-L transcripts, proteins and virus-like particles at the two-cell stage of the mouse embryo might be beneficial or deleterious to the mouse.

## Conclusions

The reconstruction of an ancestral MuERV-L sequence highlights the potential for the retroviral fossil record to illuminate ancient events and represents a unique opportunity to study ERV-L biology and reactivation, the role of MuERV-L in mouse development and potentially uncover new roles for ERVs in mammalian biology.

## Additional files

**Additional file 1: Table S1.** MuERV-L loci identified in mouse genome assemblies.

**Additional file 2.** ancML sequence in FASTA format.

**Additional file 3: Table S2.** ancML integration sites cloned from CHO gDNA.

**Additional file 4: Table S3.** Analysis for hypermutation in MuERV-L elements in the mouse genome.

## Authors' contributions

DBM performed all the experimental work and analysis. RG supervised the computational work and PDB supervised the experimental work. All authors wrote, read and approved the final manuscript.

## Author details

<sup>1</sup> Laboratory of Retrovirology and Howard Hughes Medical Institute, The Rockefeller University, New York, NY, USA. <sup>2</sup> MRC-University of Glasgow Centre for Virus Research, Glasgow, UK. <sup>3</sup> Present Address: Department of Microbiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

## Acknowledgements

We thank Dr. John V. Moran for kindly sharing the L1.3 plasmid (JM101), Theodora Hatzioannou for the plasmid expressing Fv1<sup>bbn</sup>, and Rachel Liberatore for the plasmid expressing mouse APOBEC3. We also thank all the members of the Bieniasz lab for their help and suggestions on the project.

## Competing interests

The authors declare that they have no competing interests.

## Availability of data and materials

All the datasets generated during the current study are available in the Additional files.

## Ethics approval and consent to participate

Not applicable.

## Funding

This work was supported by a grant from the National Institute of Allergy and Infectious diseases (R3764003 to PDB) and a grant from the UK Medical Research Council (MC\_UU\_12014/10 to RJG).

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 21 February 2018 Accepted: 10 April 2018

Published online: 02 May 2018

## References

- Weiss RA. The discovery of endogenous retroviruses. *Retrovirology*. 2006;3:67.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860–921.
- Emerman M, Malik HS. Paleovirology—modern consequences of ancient viruses. *PLoS Biol*. 2010;8(2):e1000301.
- Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, Pierron G, Heidmann T. Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res*. 2006;16(12):1548–56.
- Lee YN, Bieniasz PD. Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathog*. 2007;3(1):e10.
- Perez-Caballero D, Soll SJ, Bieniasz PD. Evidence for restriction of ancient primate gammaretroviruses by APOBEC3 but not TRIM5alpha proteins. *PLoS Pathog*. 2008;4(10):e1000181.
- Kaiser SM, Malik HS, Emerman M. Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein. *Science*. 2007;316(5832):1756–8.
- Goldstone DC, Yap MW, Robertson LE, Haire LF, Taylor WR, Katourakis A, Stoye JP, Taylor IA. Structural and functional analysis of prehistoric lentiviruses uncovers an ancient molecular interface. *Cell Host Microbe*. 2010;8(3):248–59.
- Soll SJ, Neil SJ, Bieniasz PD. Identification of a receptor for an extinct virus. *Proc Natl Acad Sci USA*. 2010;107(45):19496–501.
- Blanco-Melo D, Gifford RJ, Bieniasz PD. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. *Elife*. 2017;6:e22519.

11. Lee YN, Malim MH, Bieniasz PD. Hypermutation of an ancient human retrovirus by APOBEC3G. *J Virol.* 2008;82(17):8762–70.
12. Kigami D, Minami N, Takayama H, Imai H. MuERV-L is one of the earliest transcribed genes in mouse one-cell embryos. *Biol Reprod.* 2003;68(2):651–4.
13. Ribet D, Louvet-Vallee S, Harper F, de Parseval N, Dewannieux M, Heidmann O, Pierron G, Maro B, Heidmann T. Murine endogenous retrovirus MuERV-L is the progenitor of the "orphan" epsilon viruslike particles of the early mouse embryo. *J Virol.* 2008;82(3):1622–5.
14. Macfarlan TS, Gifford WD, Agarwal S, Driscoll S, Lettieri K, Wang J, Andrews SE, Franco L, Rosenfeld MG, Ren B, et al. Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. *Genes Dev.* 2011;25(6):594–607.
15. Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, Firth A, Singer O, Trono D, Pfaff SL. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature.* 2012;487(7405):57–63.
16. Costas J. Molecular characterization of the recent intragenomic spread of the murine endogenous retrovirus MuERV-L. *J Mol Evol.* 2003;56(2):181–6.
17. Benit L, Lallemand JB, Casella JF, Philippe H, Heidmann T. ERV-L elements: a family of endogenous retrovirus-like elements active throughout the evolution of mammals. *J Virol.* 1999;73(4):3301–8.
18. Lee A, Nolan A, Watson J, Tristem M. Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. *Philos Trans R Soc Lond B Biol Sci.* 2013;368(1626):20120503.
19. Cordonnier A, Casella JF, Heidmann T. Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *J Virol.* 1995;69(9):5890–7.
20. Benit L, De Parseval N, Casella JF, Callebaut I, Cordonnier A, Heidmann T. Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. *J Virol.* 1997;71(7):5652–7.
21. Magiorkinis G, Gifford RJ, Katzourakis A, De Ranter J, Belshaw R. Env-less endogenous retroviruses are genomic superspreaders. *Proc Natl Acad Sci USA.* 2012;109(19):7385–90.
22. Best S, Le Tissier P, Towers G, Stoye JP. Positional cloning of the mouse retrovirus restriction gene Fv1. *Nature.* 1996;382(6594):826–9.
23. Yan Y, Buckler-White A, Wollenberg K, Kozak CA. Origin, antiviral function and evidence for positive selection of the gammaretrovirus restriction gene Fv1 in the genus *Mus*. *Proc Natl Acad Sci USA.* 2009;106(9):3259–63.
24. Yap MW, Colbeck E, Ellis SA, Stoye JP. Evolution of the retroviral restriction gene Fv1: inhibition of non-MLV retroviruses. *PLoS Pathog.* 2014;10(3):e1003968.
25. Guallar D, Perez-Palacios R, Climent M, Martinez-Abadia I, Larraga A, Fernandez-Juan M, Vallejo C, Muniesa P, Schoorlemmer J. Expression of endogenous retroviruses is negatively regulated by the pluripotency marker Rex1/Zfp42. *Nucleic Acids Res.* 2012;40(18):8993–9007.
26. Rowe HM, Trono D. Dynamic control of endogenous retroviruses during development. *Virology.* 2011;411(2):273–87.
27. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10:421.
28. Mural RJ, Adams MD, Myers EW, Smith HO, Miklos GL, Wides R, Halpern A, Li PW, Sutton GG, Nadeau J, et al. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science.* 2002;296(5573):1661–71.
29. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature.* 2002;420(6915):520–62.
30. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, et al. The Ensembl genome database project. *Nucleic Acids Res.* 2002;30(1):38–41.
31. Swofford DL. PAUP\*. Phylogenetic analysis using parsimony (\*and other methods). Version 4. Sunderland: Sinauer Associates; 2002.
32. Lebedev YB, Belonovitch OS, Zybrowa NV, Khil PP, Kurdyukov SG, Vinogradova TV, Hunsmann G, Sverdllov ED. Differences in HERV-K LTR insertions in orthologous loci of humans and great apes. *Gene.* 2000;247(1–2):265–77.
33. Subramanian RP, Wildschutte JH, Russo C, Coffin JM. Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology.* 2011;8:90.
34. Marshall HM, Ronen K, Berry C, Llano M, Sutherland H, Saenz D, Bickmore W, Poeschla E, Bushman FD. Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS ONE.* 2007;2(12):e1340.
35. Xu X, Nagarajan H, Lewis NE, Pan S, Cai Z, Liu X, Chen W, Xie M, Wang W, Hammond S, et al. The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat Biotechnol.* 2011;29(8):735–41.
36. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32(5):1792–7.
37. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 2014;30(9):1312–3.
38. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 1997;13(5):555–6.
39. Darrriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods.* 2012;9(8):772.
40. Rose PP, Korber BT. Detecting hypermutations in viral sequences with an emphasis on G → A hypermutation. *Bioinformatics.* 2000;16(4):400–1.
41. Moran JV, DeBerardinis RJ, Kazazian HH Jr. Exon shuffling by L1 retrotransposition. *Science.* 1999;283(5407):1530–4.
42. Kinomoto M, Kanno T, Shimura M, Ishizaka Y, Kojima A, Kurata T, Sata T, Tokunaga K. All APOBEC3 family proteins differentially inhibit LINE-1 retrotransposition. *Nucleic Acids Res.* 2007;35(9):2955–64.
43. Mariani R, Chen D, Schrofelbauer B, Navarro F, Konig R, Bollman B, Munk C, Nymark-McMahon H, Landau NR. Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. *Cell.* 2003;114(1):21–31.
44. Hatziaioannou T, Cowan S, Bieniasz PD. Capsid-dependent and -independent postentry restriction of primate lentivirus tropism in rodent cells. *J Virol.* 2004;78(2):1006–11.
45. Soneoka Y, Cannon PM, Ramsdale EE, Griffiths JC, Romano G, Kingsman SM, Kingsman AJ. A transient three-plasmid expression system for the production of high titer retroviral vectors. *Nucleic Acids Res.* 1995;23(4):628–33.
46. Kent WJ. BLAT—the BLAST-like alignment tool. *Genome Res.* 2002;12(4):656–64.
47. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002;12(6):996–1006.
48. Brady T, Lee YN, Ronen K, Malani N, Berry CC, Bieniasz PD, Bushman FD. Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev.* 2009;23(5):633–42.
49. Soh YQ, Alfoldi J, Pyntikova T, Brown LG, Graves T, Minx PJ, Fulton RS, Kremitzki C, Koutseva N, Mueller JL, et al. Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell.* 2014;159(4):800–13.
50. Qin C, Wang Z, Shang J, Bekkari K, Liu R, Pacchione S, McNulty KA, Ng A, Barnum JE, Storer RD. Intracisternal A particle genes: distribution in the mouse genome, active subtypes, and potential roles as species-specific mediators of susceptibility to cancer. *Mol Carcinog.* 2010;49(1):54–67.
51. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr. High frequency retrotransposition in cultured mammalian cells. *Cell.* 1996;87(5):917–27.
52. Lewis NE, Liu X, Li Y, Nagarajan H, Yerganian G, O'Brien E, Bordbar A, Roth AM, Rosenbloom J, Bian C, et al. Genomic landscapes of Chinese hamster ovary cell lines as revealed by the *Cricetulus griseus* draft genome. *Nat Biotechnol.* 2013;31(8):759–65.
53. Repaske R, Steele PE, O'Neill RR, Rabson AB, Martin MA. Nucleotide sequence of a full-length human endogenous retroviral segment. *J Virol.* 1985;54(3):764–72.
54. Coffin JM, Hughes SH, Varmus H. *Retroviruses*. Plainview, NY: Cold Spring Harbor Laboratory Press; 1997.
55. Champoux JJ, Schultz SJ. Ribonuclease H: properties, substrate specificity and roles in retroviral reverse transcription. *FEBS J.* 2009;276(6):1506–16.
56. Schoggins JW, Wilson SJ, Panis M, Murphy MY, Jones CT, Bieniasz P, Rice CM. A diverse range of gene products are effectors of the type I interferon antiviral response. *Nature.* 2011;472(7344):481–5.
57. Zwarthoff EC, Bosveld IJ, Vonk WP, Trapman J. Constitutive expression of a murine interferon alpha gene in hamster cells and characterization of its protein product. *J Gen Virol.* 1985;66(Pt 4):685–91.
58. Van Heuvel M, Bosveld IJ, Mooren AA, Trapman J, Zwarthoff EC. Properties of natural and hybrid murine alpha interferons. *J Gen Virol.* 1986;67(Pt 10):2215–22.



59. Trapman J, van Heuvel M, de Jonge P, Bosveld IJ, Klaassen P, Zwarthoff EC. Structure-function analysis of mouse interferon alpha species: MulFN-alpha 10, a subspecies with low antiviral activity. *J Gen Virol*. 1988;69(Pt 1):67–75.
60. Van Heuvel M, Bosveld IJ, Klaassen P, Zwarthoff EC, Trapman J. Structure-function analysis of murine interferon-alpha: antiviral properties of novel hybrid interferons. *J Interferon Res*. 1988;8(1):5–14.
61. van Heuvel M, Govaert-Siemerink M, Bosveld IJ, Zwarthoff EC, Trapman J. Interferon-alpha-(IFN) producing CHO cell lines are resistant to the antiproliferative activity of IFN: a correlation with gene expression. *J Cell Biochem*. 1988;38(4):269–78.
62. Wu X, Dao Thi VL, Huang Y, Billerbeck E, Saha D, Hoffmann HH, Wang Y, Silva LAV, Sarbanes S, Sun T, et al. Intrinsic immunity shapes viral resistance of stem cells. *Cell*. 2018;172(3):423–38.
63. Bock M, Bishop KN, Towers G, Stoye JP. Use of a transient assay for studying the genetic determinants of Fv1 restriction. *J Virol*. 2000;74(16):7422–30.
64. Rehwinkel J. Mouse knockout models for HIV-1 restriction factors. *Cell Mol Life Sci*. 2014;71(19):3749–66.
65. Lovsin N, Peterlin BM. APOBEC3 proteins inhibit LINE-1 retrotransposition in the absence of ORF1p binding. *Ann NY Acad Sci*. 2009;1178:268–75.
66. Muckenfuss H, Hamdorf M, Held U, Perkovic M, Lower J, Cichutek K, Flory E, Schumann GG, Munk C. APOBEC3 proteins inhibit human LINE-1 retrotransposition. *J Biol Chem*. 2006;281(31):22161–72.
67. Stenglein MD, Harris RS. APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. *J Biol Chem*. 2006;281(25):16837–41.
68. Xie D, Chen CC, Ptaszek LM, Xiao S, Cao X, Fang F, Ng HH, Lewin HA, Cowan C, Zhong S. Rewirable gene regulatory networks in the pre-implantation embryonic development of three mammalian species. *Genome Res*. 2010;20(6):804–15.
69. MacMillan AL, Kohli RM, Ross SR. APOBEC3 inhibition of mouse mammary tumor virus infection: the role of cytidine deamination versus inhibition of reverse transcription. *J Virol*. 2013;87(9):4808–17.
70. Lavielle C, Cornelis G, Dupressoir A, Esnault C, Heidmann O, Vernochet C, Heidmann T. Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philos Trans R Soc Lond B Biol Sci*. 2013;368(1626):20120507.
71. Wang J, Xie G, Singh M, Ghanbarian AT, Rasko T, Szvetnik A, Cai H, Besser D, Prigione A, Fuchs NV, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*. 2014;516(7531):405–9.
72. Lu X, Sachs F, Ramsay L, Jacques PE, Goke J, Bourque G, Ng HH. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol*. 2014;21(4):423–5.
73. Armezzani A, Varela M, Spencer TE, Palmarini M, Arnaud F. "Menage a trois": the evolutionary interplay between JSRV, enJSRVs and domestic sheep. *Viruses*. 2014;6(12):4926–45.
74. Huang Y, Kim JK, Do DV, Lee C, Penfold CA, Zyllicz JJ, Marioni JC, Hackett JA, Surani MA. Stella modulates transcriptional and endogenous retrovirus programs during maternal-to-zygotic transition. *Elife*. 2017;6:e22345.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

