Genome-wide inactivation of porcine endogenous retroviruses (PERVs)

Luhan Yang[1,2,3,†,*], Marc Güell[1,2,3, †], Dong Niu[1,4, †], Haydy George[1, †], Emal Lesha[1], Dennis Grishin[1], John Aach[1], Ellen Shrock[1], Weihong Xu[6], Jürgen Poci[1], Rebeca Cortazio[1], Robert A Wilkinson[5], Jay A. Fishman[5], George Church[1,2,3,*]

Affiliations:

1. Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA

2. Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, Massachusetts, USA

3. eGenesis Biosciences, Boston, MA 02115, USA

4. College of Animal Sciences, Zhejiang University, Hangzhou 310058, China

5. Transplant Infectious Disease & Compromised Host Program, Massachusetts General Hospital, Boston, MA 02115, USA

6. Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

*, Correspondence should be addressed to gchurch@genetics.med.harvard.edu;

luhan.yang@egenesisbio.com

†. These authors contributed equally to this work

Abstract:

The shortage of organs for transplantation is a major barrier to the treatment of organ failure. While porcine organs are considered promising, their use has been checked by concerns about transmission of porcine endogenous retroviruses (PERVs) to humans. Here, we describe the eradication of all PERVs in a porcine kidney epithelial cell line (PK15). We first determined the PK15 PERV copy number to be 62. Using CRISPR-Cas9, we disrupted all 62 copies of the PERV *pol* gene and demonstrated a > 1000-fold reduction in PERV transmission to human cells using our engineered cells. Our study shows that CRISPR-Cas9 multiplexability can be as high as 62 and demonstrates the possibility that PERVs can be inactivated for clinical application of porcine-to-human xenotransplantation.

One sentence summary:

CRISPR-cas9 can be used to inactivate all 62 copies of a porcine endogenous retrovirus from a swine genome (PERVs).

Pig genomes contain from a few to several dozen copies of PERV elements (*1*). Unlike other zoonotic pathogens, PERVs cannot be eliminated by biosecure breeding (*2*). Thus, strategies for reducing the risk of PERV transmission to humans have included small interfering RNAs (RNAi), vaccines (*3*–*5*), and PERV elimination using zinc finger nucleases (*6*) and TAL effector nucleases (*7*) , but these have had limited success. Here we report the successful use of the CRISPR-Cas9

RNA-guided nuclease system (*8*) (*9*, *10*) to inactivate all copies of the PERV *pol* gene and show that this reduces PERV infectivity of human cells by 1000-fold.

To design Cas9 guide RNAs (gRNAs) that specifically target PERVs, we analyzed the sequences of publically available PERVs and other endogenous retroviruses in pigs (Methods). We identified a distinct clade of PERV elements (Fig. 1A) and determined there to be 62 copies of PERVs in PK15 cells using droplet digital PCR (Fig. 1B). We then designed two Cas9 guide RNAs (gRNAs) that targeted the highly conserved catalytic center (*11*) of the *pol* gene on PERVs(Fig. 1C, Fig. S1). The *pol* gene product functions as a reverse transcriptase (RT) and is thus essential for viral replication and infection. We determined that these gRNAs targeted all PERVs but no other endogenous retrovirus or other sequences in the pig genome (Methods).

Initial experiments showed inefficient PERV editing when Cas9 and the gRNAs were transiently transfected (Fig. S2). Thus we used a PiggyBac transposon (*12*) system to deliver a doxycycline-inducible Cas9 and the two gRNAs into the genome of PK15 cells (Fig. S2-3). Continuous induction of Cas9 led to increased targeting frequency of the PERVs (Fig. S5), with a maximum targeting frequency of 37% (~23 PERV copies per genome) observed on day 17 (Fig. S5). Neither higher concentrations of doxycycline or prolonged incubation increased targeting efficiency (Fig. S4,5), possibly due to the toxicity of non-specific DNA damage by CRISPR-Cas9. Similar trends were observed when Cas9 was delivered using lentiviral constructs (Fig. S6). We then genotyped the cell lines that exhibited maximal PERV targeting efficiencies. We observed 455 different insertion and deletion (indel) events centered at the two gRNA target sites (Fig. 2B). Indel sizes ranged from 1 to 148bp, 80% of which were small deletions (<9bp). We validated the initial deep sequencing results with Sanger Sequencing (Fig. S7).

We next sorted single cells from PK15 cells with high PERV targeting efficiency using flow cytometry and genotyped the *pol* locus of the resulting clones via deep sequencing (*13*, *14*). A repeatable bimodal (Fig. 2A, S8-9) distribution was observed with ~10% of the clones exhibiting high levels of PERV disruption (97%-100%), and the remaining clones exhibiting low levels of editing (<10%). We then examined individual indel events in the genomes of these clones (Fig. 2B, Fig. S10-11). For the highly edited clones (Clone 20, 100%; Clone 15, 100%; Clone 29, 100%; Clone 38, 97.37%), there were only 16-20 unique indel patterns in each clone (Fig. 2B, S11). In addition, there was much higher degree shared of indels within the clones than across the clones (Fig. S25), suggesting that gene conversion might have operated during editing to spread previously mutated PERV copies to wild-type PERVs cut by Cas9 (Fig. 2B, Fig. S25). Mathematical modeling of DNA repair during PERV elimination (Fig. S26), and analysis of expression data (Fig. S22-24) supported this hypothesis and suggested that highly edited clones came from cells in which Cas9 and the gRNAs were highly expressed.

Next, we examined whether unexpected genomic rearrangements had occurred as a result of the multiplexed genome editing. Karyotyping of individual modified clones (Fig. S12-S14) indicated that there were no observable genomic rearrangements. We also examined 11 independent genomic loci with at most 2bp mismatches to each of the intended gRNA targets and observed no non-specific mutations (Fig. S27 and methods). This suggests that our multiplexed Cas9-based genome engineering strategy did not cause catastrophic genomic instability.

Last, we examined whether our disruption of all copies of PERV *pol* in the pig genome could eliminate *in vitro* transmission of PERVs from pig to human cells. We could not detect RT activity

in the cell culture supernatant of the highly modified PK15 clones (Fig. S15), suggesting that modified cells can only produce minimal amounts of PERV particles. We then co-cultured WT and highly modified PK15 cells with HEK 293 cells to test directly for transmission of PERV DNA to the human cells (*15*). After co-culturing PK15 WT and HEK 293 cells for 5 days and 7 days (Fig. 3A, S16-17), we detected PERV *pol, gag,* and *env* sequences in the HEK 293 cells (Fig. 3A). We estimated the frequency of PERV infection to be approximately 1000 PERVs/ 1000 human cells (Fig. 3B). However, PK15 clones with > 97% PERV *pol* targeting exhibited up to 1000-fold reduction of PERV infection, similar to background levels (Fig. 3C). We validated these results with PCR amplification of serial dilutions of HEK293 cells that had a history of contact with PK15 clones (Fig. 3D, S18-21). We could consistently detect PERVs in single HEK293 cells isolated from the population co-cultured with minimally modified Clone 40, but we could not distinctly detect PERVs in 100 human cells from the population co-cultured with highly modified Clone 20. Thus, we concluded that the PERV infectivity of the engineered PK15 cells had been reduced by up to 1000 fold.

In summary, we successfully targeted the 62 copies of PERV *pol* in PK15 cells and demonstrated greatly reduced *in vitro* transmission of PERVs to human cells. While *in vivo* PERV transmission to humans has not been demonstrated (*16*, *17*), PERVs are still considered risky (*18*, *19*) and this strategy could completely eliminate this risk. As no porcine embryonic stem cells exist, this system will need to be recapitulated in primary porcine cells and cloned into animals using somatic cell nuclear transfer procedure. Moreover, we achieved simultaneous Cas9 targeting of 62 loci in single pig cells without salient genomic rearrangement. To our knowledge, the maximum number of genomic sites previously reported to be simultaneously edited has been six (*20*). Our methods thus open the possibility of editing other repetitive regions of biological significance.

References

1. D. Lee *et al.*, Rapid determination of perv copy number from porcine genomic DNA by real-time polymerase chain reaction. *Anim. Biotechnol.* **22**, 175–80 (2011).
2. H.-J. Schuurman, The International Xenotransplantation Association consensus statement on conditions for undertaking clinical trials of porcine islet products in type 1 diabetes--chapter 2: Source pigs. *Xenotransplantation*. **16**, 215–22.
3. J. Ramsoondar *et al.*, Production of transgenic pigs that express porcine endogenous retrovirus small interfering RNAs. *Xenotransplantation*. **16**, 164–80.
4. M. Semaan, D. Kaulitz, B. Petersen, H. Niemann, J. Denner, Long-term effects of PERV-specific RNA interference in transgenic pigs. *Xenotransplantation*. **19**, 112–21.
5. U. Fiebig, O. Stephan, R. Kurth, J. Denner, Neutralizing antibodies against conserved domains of p15E of porcine endogenous retroviruses: basis for a vaccine for xenotransplantation? *Virology*. **307**, 406–13 (2003).
6. M. Semaan, D. Ivanusic, J. Denner, Cytotoxic Effects during Knock Out of Multiple Porcine Endogenous Retrovirus (PERV) Sequences in the Pig Genome by Zinc Finger Nucleases (ZFN). *PLoS One*. **10**, e0122059 (2015).
7. D. Dunn, M. DaCosta, M. Harris, R. Idriss, A. O'Brien, Genetic Modification of Porcine Endogenous Retrovirus (PERV) Sequences in Cultured Pig Cells as a Model for Decreasing Infectious Risk in Xenotransplantation. *FASEB J.* **29**, LB761– (2015).
8. M. Jinek *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*. **337**, 816–21 (2012).

9. P. Mali *et al.*, RNA-guided human genome engineering via Cas9. *Science*. **339**, 823–6 (2013).

10. L. Cong *et al.*, Multiplex genome engineering using CRISPR/Cas systems. *Science*. **339**, 819–23 (2013).

11. S. G. Sarafianos *et al.*, Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition. *J. Mol. Biol.* **385**, 693–713 (2009).

12. M. H. Wilson, C. J. Coates, A. L. George, PiggyBac transposon-mediated gene transfer in human cells. *Mol. Ther.* **15**, 139–45 (2007).

13. L. Yang *et al.*, Optimization of scarless human stem cell genome editing. *Nucleic Acids Res.* **41**, 9049–61 (2013).

14. M. Güell, L. Yang, G. M. Church, Genome editing assessment using CRISPR Genome Analyzer (CRISPR-GA). *Bioinformatics*. **30**, 2968–2970 (2014).

15. C. Patience, Y. Takeuchi, R. A. Weiss, Infection of human cells by an endogenous retrovirus of pigs. *Nat. Med.* **3**, 282–6 (1997).

16. W. Heneine *et al.*, No evidence of infection with porcine endogenous retrovirus in recipients of porcine islet-cell xenografts. *Lancet*. **352**, 695–9 (1998).

17. J. H. Dinsmore, C. Manhart, R. Raineri, D. B. Jacoby, A. Moore, No evidence for infection of human cells with porcine endogenous retrovirus (PERV) after exposure to porcine fetal neuronal cells. *Transplantation*. **70**, 1382–9 (2000).

18. D. Butler, Last chance to stop and think on risks of xenotransplants. *Nature*. **391**, 320–4 (1998).

19. Xenotransplantation: Science, Ethics, and Public Policy. *ILAR J.* **38**, 49–51 (1997).

20. H. Wang *et al.*, One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell*. **153**, 910–8 (2013).

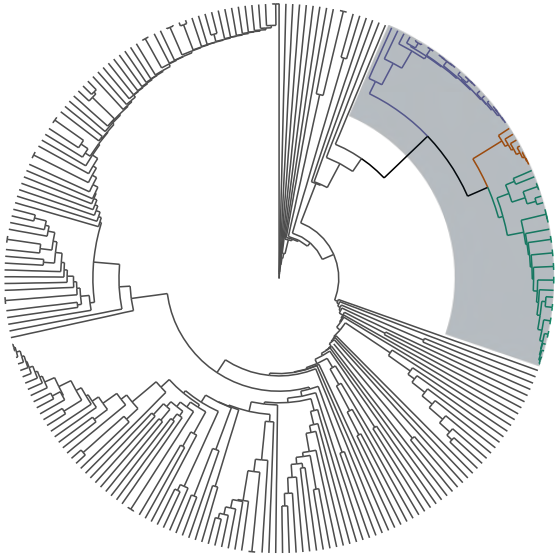
Acknowledgments: We thank Zach Herbert and Mahesh Vangala at the Dana Farber MBC for assistance with RNA analysis, Yichen Shen and Oscar Castanon for volunteering in the lab research, Samuel Broder for scientific advices. MG is funded by a Human Frontiers Science Program Long Term fellowship. This study was funded by NIH grant P50 HG005550.

Figure 1| CRISPR-Cas9 gRNAs were designed to specifically target the *pol* gene in 62 copies of PERVs in PK15 cells. (**A)** Phylogenetic tree representing endogenous retroviruses present in the pig genome. PERVs are highlighted in blue. (**B**) Copy number determination of PERVs in PK15 cells *via* digital droplet PCR. The copy number of *pol* elements was estimated to be 62 using three independent reference genes: *ACTB*, *GAPDH*, and *EB2*. N=3, mean +/- SEM. (**C)** We designed two CRISPR-Cas9 gRNAs to target the catalytic region of the PERV *pol* gene. The two gRNA targeting sequences are shown below a schematic of PERV gene structure. Their PAM sequences are highlighted in red.

Figure 2| Clonal PK15 cells with inactivation of all copies of PREV *pol* genes after Cas9 treatment. **(A)** A bimodal distribution of *pol* targeting efficiencies was observed among the single-cell-derived PK15 clones after 17 days of Cas9 induction. 45/50 exhibited <16% targeting efficiency; 5/50 clones exhibited >93% targeting efficiency. **(B)** PK15 haplotypes at PERV *pol* loci after CRISPR-Cas9 treatment. In red, indel events in the PERV *pol* sequence are represented. Shades of purple indicate endogenous PERVs.

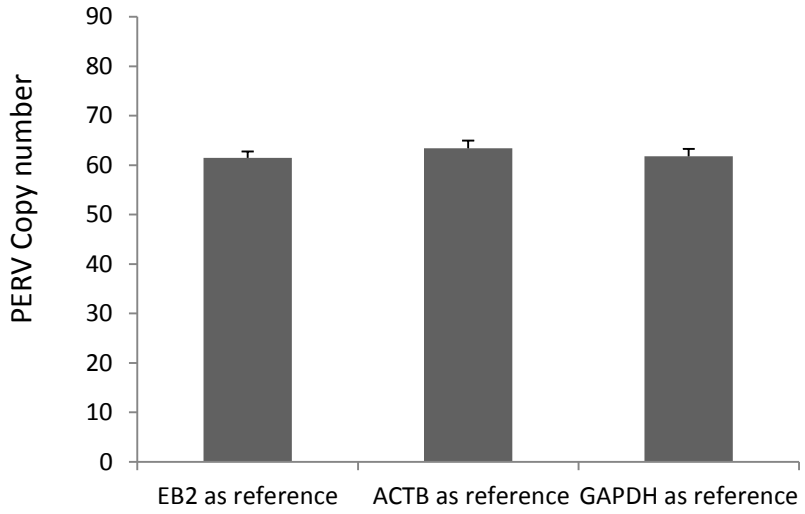
Figure 3| **(A)** Detection of PERV *pol, gag*, and *env* DNA in the genomes of HEK-293-GFP cells after co-culturing with PK15 cells for 5 days and 7 days (293G5D and 393G7D, respectively). A pig GGTA1 primer set was used to detect pig cell contamination of the purified human cells. **(B)** qPCR quantification of the number of PERV elements in 1000 293G cells derived from a population co-cultured with wild type PK15 cells using specific primer sets. (N=3, mean+/-SEM) **(C)** qPCR quantification of the number of PERV elements in PK15 Clones 15, 20, 29, and 38, with high levels of PERV *pol* modification, and minimally modified Clones 40 and 41. (N=3, mean+/-SEM) **(D)** Results of PCR on PERV *pol* on genomic DNA from various numbers of HEK 293-GFP cells (0.1, 1, 10,and 100) isolated from populations previously cultured with highly modified PK15 Clone 20 and minimally modified Clone 40. See Fig. S18-21 for a full panel of PCR reactions.

A

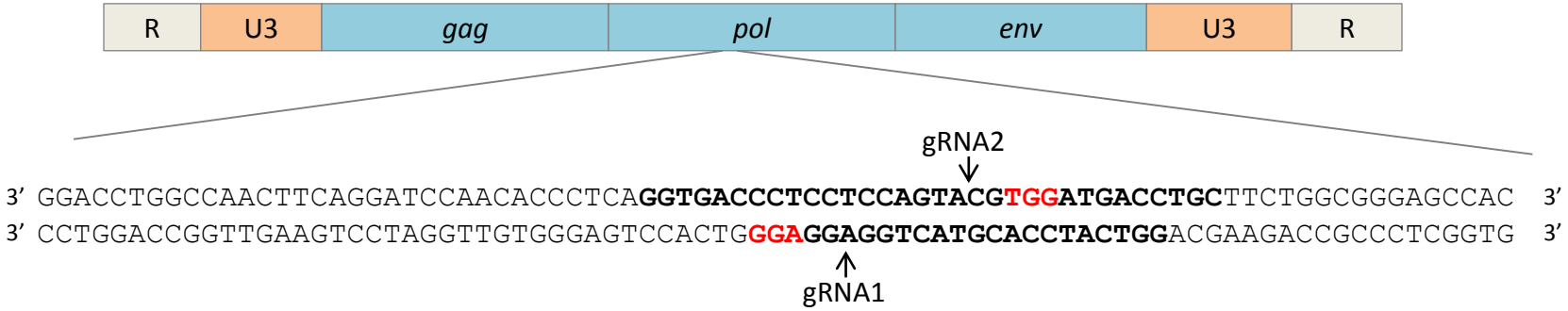


PERVs

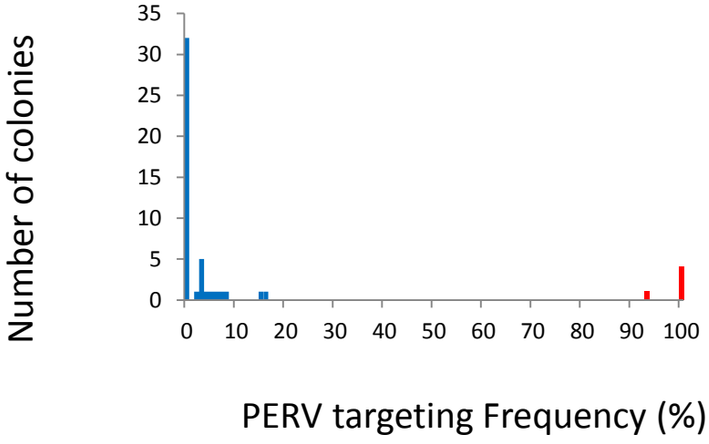
B


PERV Copy number
90
80
70
60
50
40
30
20
10
0
EB2 as reference
ACTB as reference
GAPDH as reference

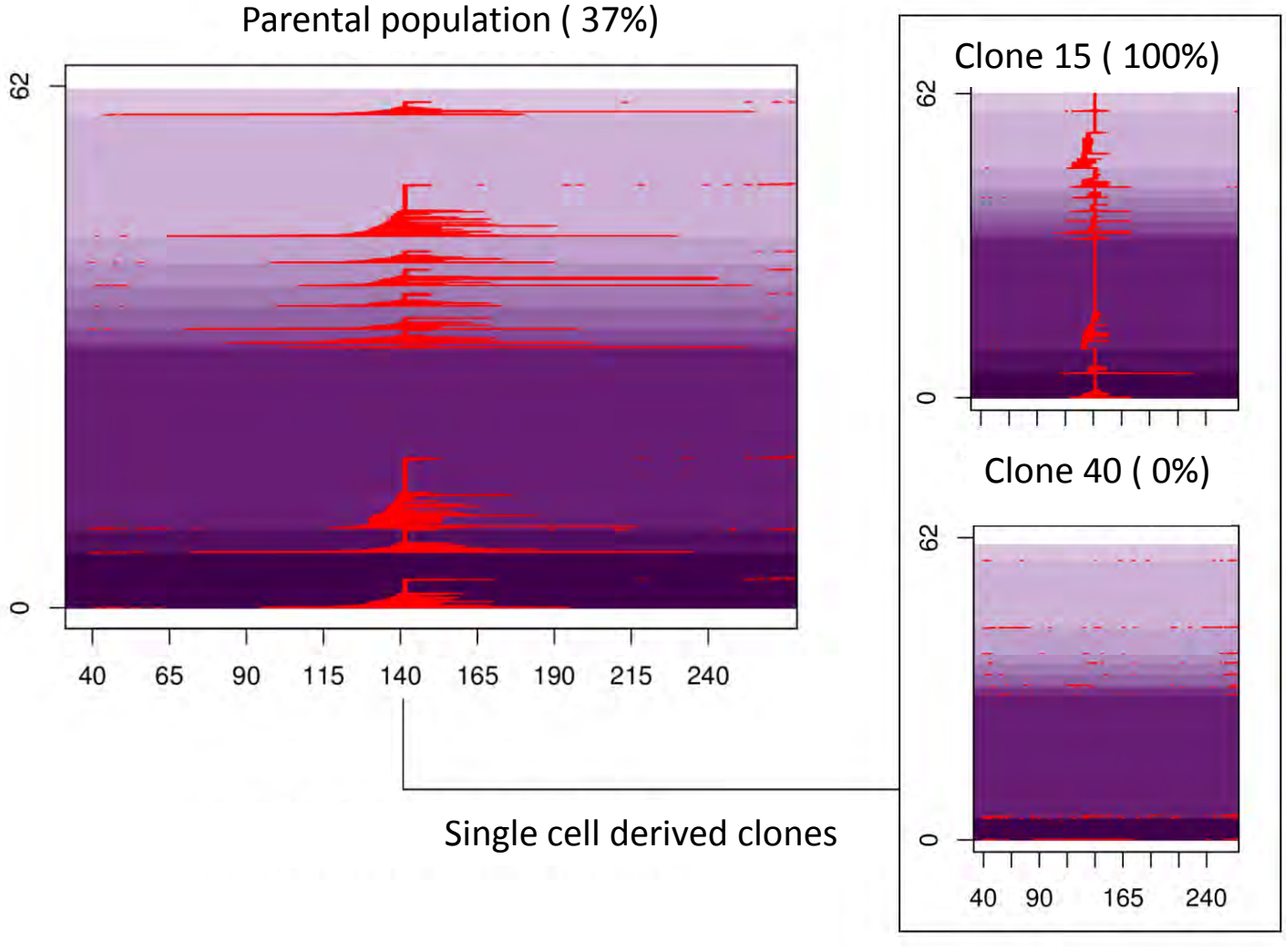
C

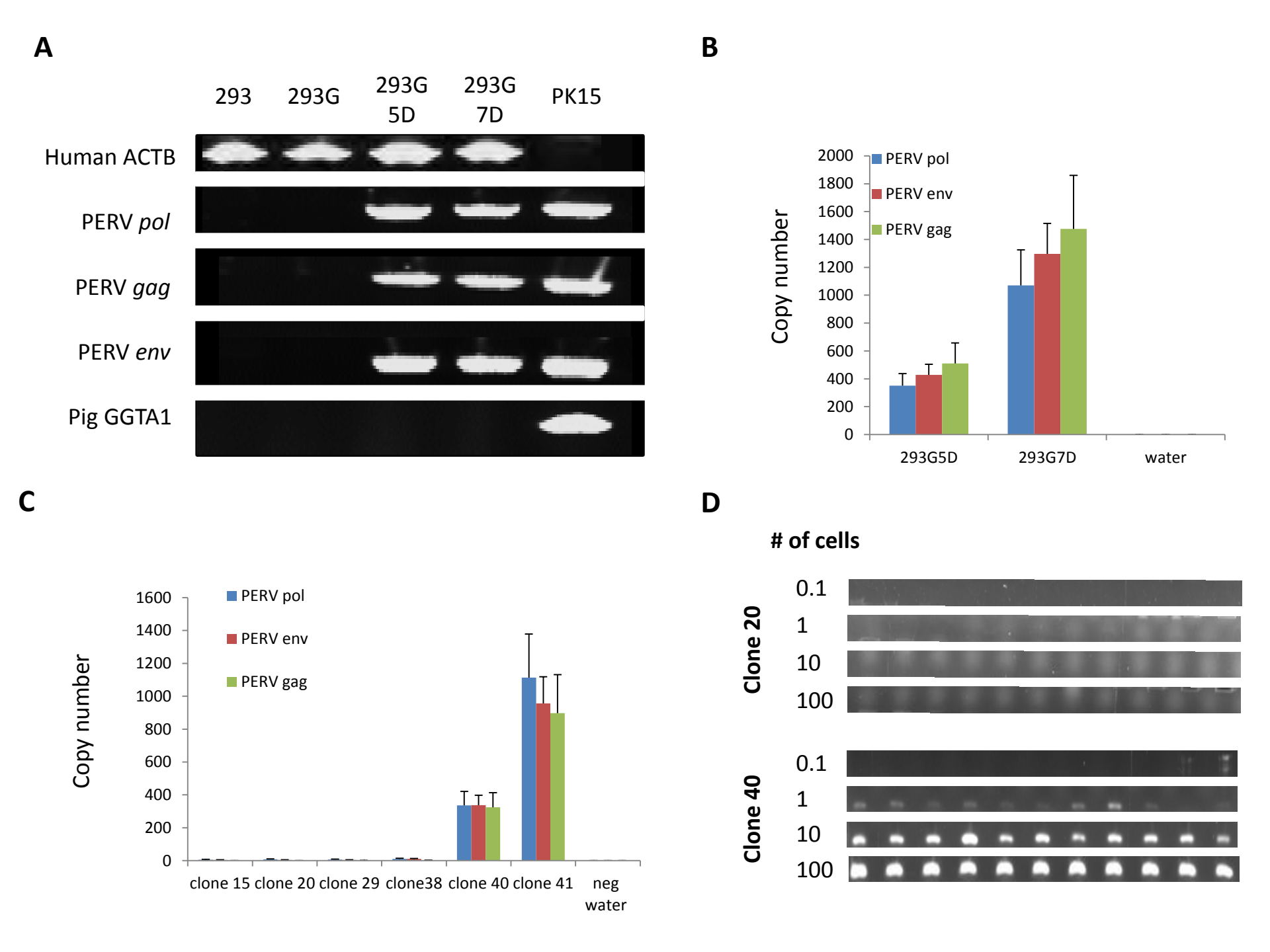

R
U3
gag
pol
env
U3
R
gRNA2
3' GGACCTGGCCAACTTCAGGATCCAACACCCTCAGGTGACCCTCCTCCAGTACGTGGATGACCTGCTTCTGGCGGGAGCCAC 3'
3' CCTGGACCGGTTGAAGTCCTAGGTTGTGGGAGTCCACTGGGAGGAGGTCATGCACCTACTGGACGAAGACCGCCCTCGGTG 3'
gRNA1

A


Number of colonies
35
30
25
20
15
10
5
0
0 10 20 30 40 50 60 70 80 90 100
PERV targeting Frequency (%)

B


Parental population (37%)
62
0
40 65 90 115 140 165 190 215 240
Clone 15 (100%)
62
0
Clone 40 (0%)
62
0
40 90 165 240
Single cell derived clones


A
293
293G
293G 5D
293G 7D
PK15
Human ACTB
PERV pol
PERV gag
PERV env
Pig GGTA1
B
Copy number
2000
1800
1600
1400
1200
1000
800
600
400
200
0
PERV pol
PERV env
PERV gag
293G5D
293G7D
water
C
Copy number
1600
1400
1200
1000
800
600
400
200
0
PERV pol
PERV env
PERV gag
clone 15
clone 20
clone 29
clone38
clone 40
clone 41
neg water
D
# of cells
Clone 20
0.1
1
10
100
Clone 40
0.1
1
10
100

Supplementary Information

Genome-wide inactivation of porcine endogenous retroviruses (PERVs)

Luhan Yang[1,2,3,†,*], Marc Güell[1,2,3,†], Dong Niu[1,4,†], Haydy George[1,†], Emal Lesha[1], Dennis Grishin[1], John Aach[1], Ellen Shrock[1], Weihong Xu[6], Jürgen Poci[1], Rebeca Cortazio[1], Robert A Wilkinson[5], Jay A. Fishman[5], George Church[1,2,3,*]

Affiliations:

1. Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA

2. Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, Massachusetts, USA

3. eGenesis Biosciences, Boston, MA 02115, USA

4. College of Animal Sciences, Zhejiang University, Hangzhou 310058, China

5. Transplant Infectious Disease & Compromised Host Program, Massachusetts General Hospital, Boston, MA 02115, USA

6. Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

*, Correspondence should be addressed to gchurch@genetics.med.harvard.edu;

luhan.yang@egenesisbio.com

†. These authors contributed equally to this work

Figure S1	PERV *pol* consensus sequence and gRNA design
Figure S2	Schematic picture of CRISPR/Cas9 construct targeting PERV
Figure S3	Measurement of Cas9-gRNAs activity
Figure S4	Optimization of DOX concentration to induce Cas9 expression for PERV targeting
Figure S5	Time series measurement of Piggybac-Cas9/gRNAs PERV targeting efficiencies
Figure S6	Time series measurement of Lenti-Cas9/2gRNAs PERV targeting efficiency
Figure S7	Sanger sequencing validation of PERV targeting efficiency and indel pattern~~ing~~
Figure S8	We repeated the gene editing experiment
Figure S9	PERV pol targeting efficiency of single cells
Figure S10	Phylogeny of PERV haplotypes
Figure S11	Distribution of the *pol* gene targeting pattern
Figure S12	Karyotype analysis of highly and lowly modified PK15 clones
Figure S13	Summary of karyotype analysis of PK15 clones
Figure S14	Karyotype nomenclature
Figure S15	Detection of PERV reverse transcriptase activity
Figure S16	Experimental design to detect the transmission of PERVs to human cells.
Figure S17	Quality control of the purified HEK293-GFP cells by FACS
Figure S18	Detection of pig cell contamination in HEK293 cells using pig GGTA1primers
Figure S19	Detection of PERV DNA elements in HEK293 cells using PERV *pol* primers
Figure S20	Detection of PERV DNA elements in HEK293 cells using PERV *env* primers
Figure S21	Detection of PERV DNA elements in HEK293 cells using PERV *gag* primers
Figure S22	Cas9/2gRNAs expression levels in highly and lowly modified clones
Figure S23	Principle component analysis of highly and lowly modified PK15 clones
Figure S24	Gene set enrichment analysis
Figure S25	Indels composition analysis and comparison between highly modified clones.
Figure S26	Markov model analysis of DNA repair processes leading to Cas9 elimination of active PERV elements.

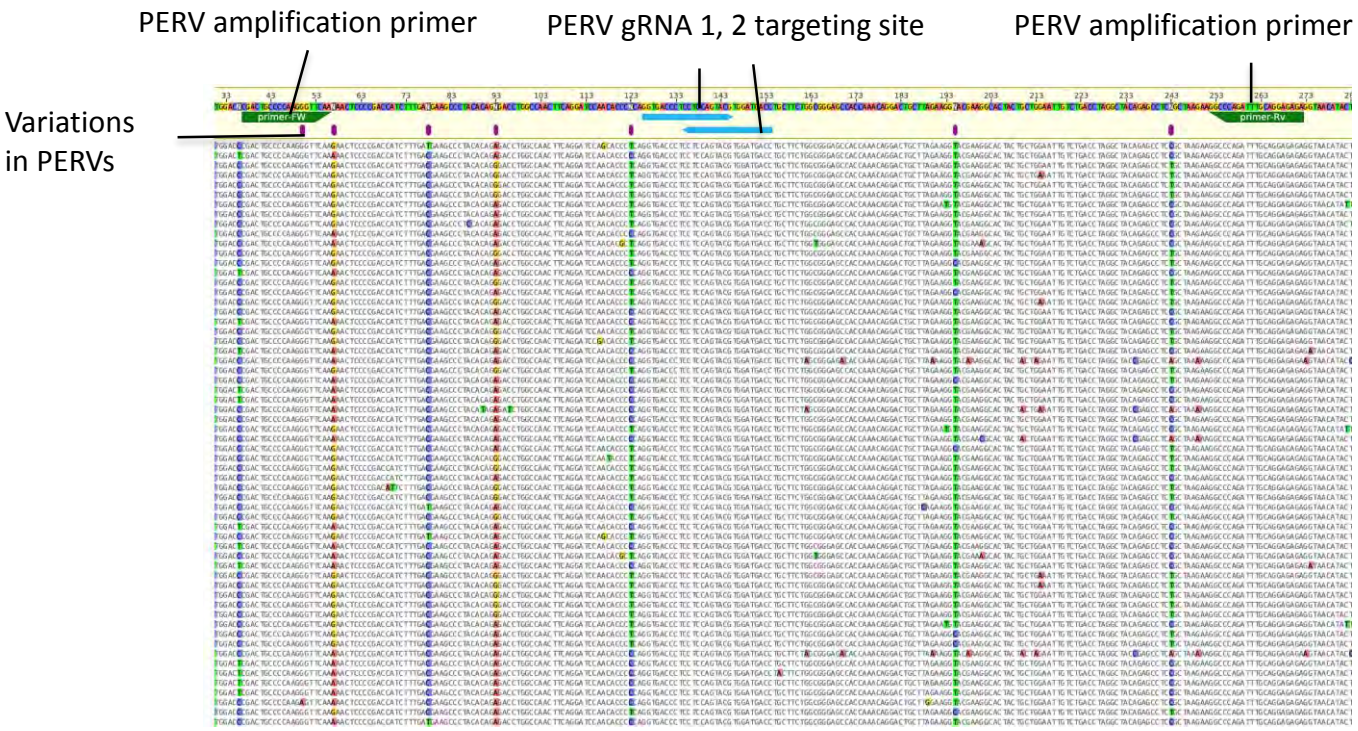


Figure S1: PERV *pol* consensus sequence and gRNA design

The consensus sequence of the PERV *pol* catalytic center is shown above. Below, PERV sequences extracted from GenBank and from a recently assembled pig genome (*1*) are listed. In blue, the gRNAs designed to target PERV *pol* are annotated; in green, the primers used to genotype the *pol* loci are annotated. Purple ovals mark the positions at which we observed variation in the sequences of the PERVs in PK15 cells; this variation was analyzed and used to generate the Phylogeny of PERV haplotypes (Fig. S8).

Figure S2: Schematic picture of CRISPR/Cas9 construct targeting PERV

Piggybac-Cas9/2gRNAs construct. Cas9 and 2 gRNAs are carried in the same vector. Cas9 expression was induced by addition of doxycycline.

.

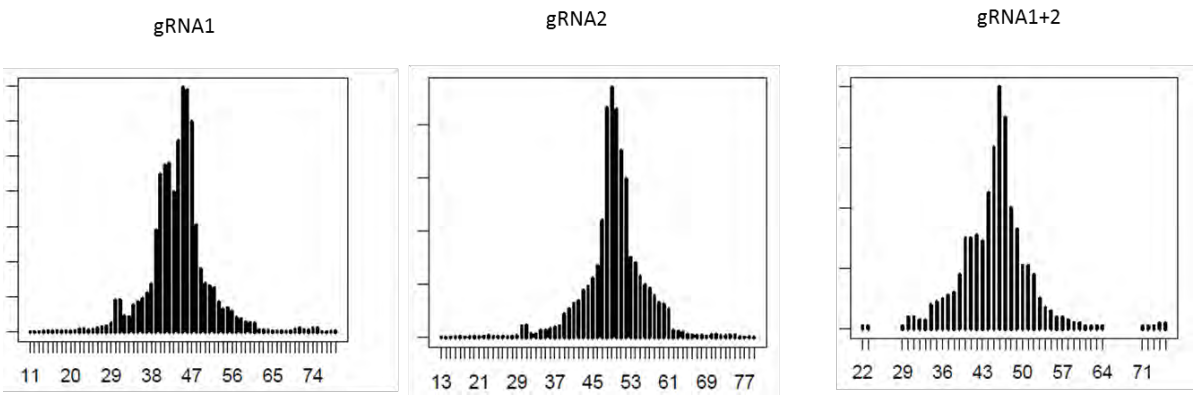


Figure S3: Measurement of of Cas9/2gRNAs activity

We tested the genomic DNA cutting activities of PERV gRNA1 and gRNA2 individually by establishing porcine fallopian tube endothelium cell (FTEC) lines integrated with individual CRISPR/gRNA constructs. Analysis of deep sequencing data revealed cutting efficiency to be 2% for gRNA1 (left), 3% for gRNA2 (middle) and 3% for gRNA1+2 (right) at individual designated cutting sites 3 days after integration.

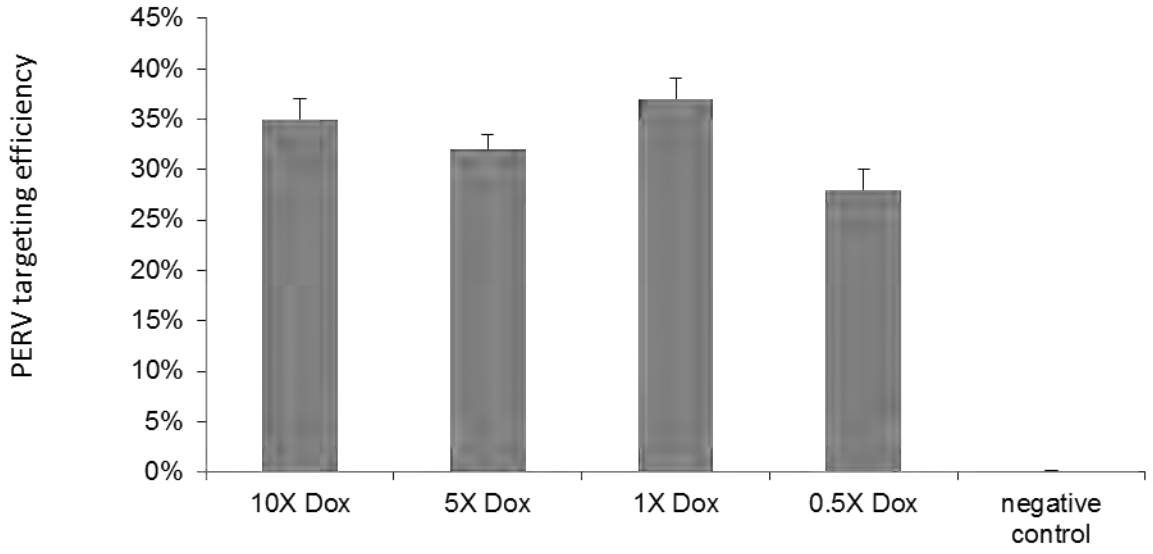


Figure S4: Optimization of DOX concentration to induce Cas9 expression for PERV targeting

We induced Cas9 expression in PK15 PiggyBac-Cas9 with different concentrations of doxycycline (DOX) and examined PERV targeting efficiency 17 days after DOX induction. We did not observe a difference in PERV targeting efficiency with 1X DOX induction as compared with 10X and 5X DOX concentrations.

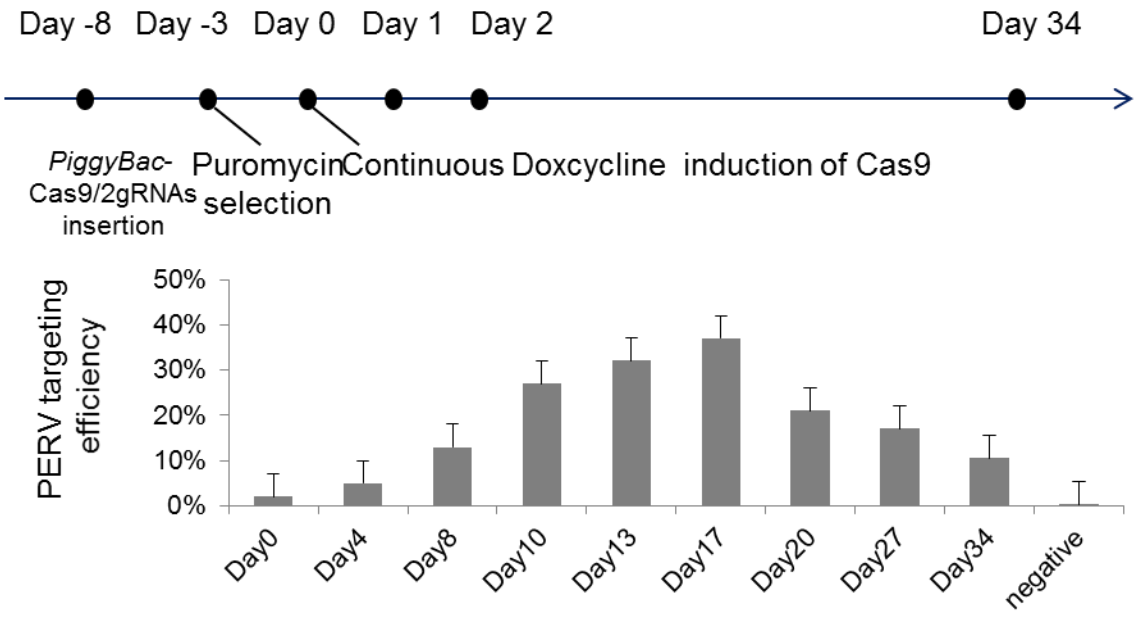


Figure S5: Time series measurement of Piggybac-Cas9/gRNAs PERV targeting efficiencies

Experimental timeline. Piggybac-Cas9/2gRNAs were integrated into PK15 cell lines and the day when puromycin selection was completed is designated as Day 0. Thereafter, the PK15-piggybac-Cas9/2gRNAs cell culture was supplemented with 1ng/ml doxycycline to induce the Cas9 expression. Cells were given fresh medium containing doxycycline every two days.

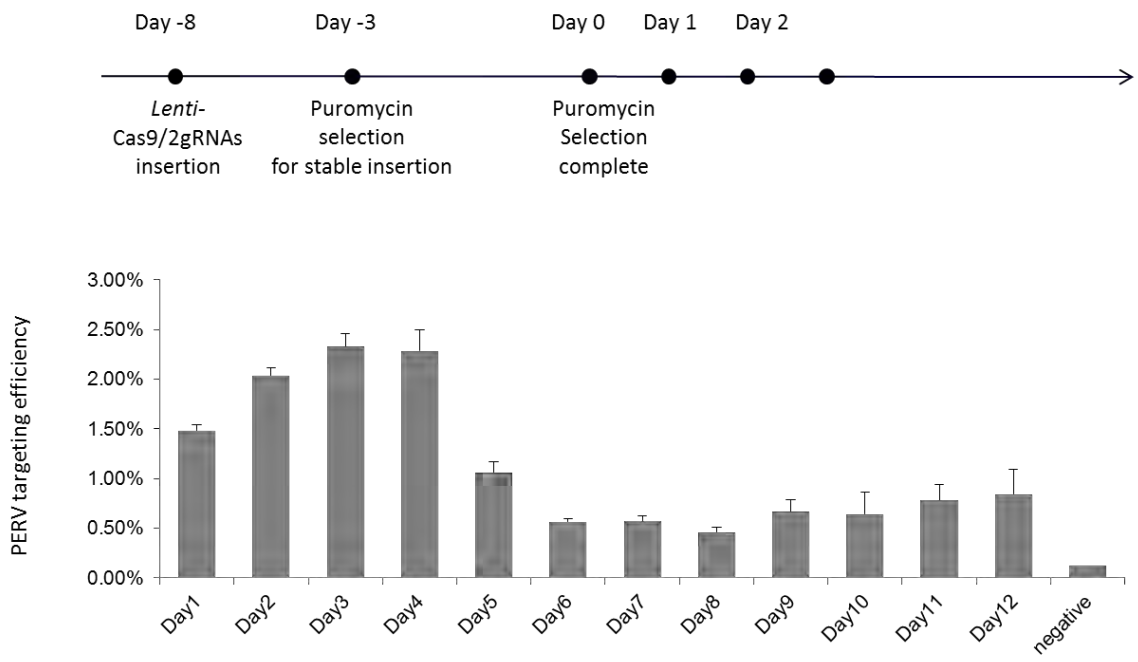


Figure S6: Time series measurement of Lenti-Cas9/gRNAs PERV targeting efficiencies

Upper panel: Timeline of the lenti-Cas9/2gRNAs targeting experiment. We integrated lenti/Cas9/2gRNAs into PK15 cell lines and designated the day when puromycin selection was completed as Day 0. The Cas9 expression is constitutively driven by pEF-1a; Bottom panel: the PERV targeting efficiency as measured by deep sequencing. n=2

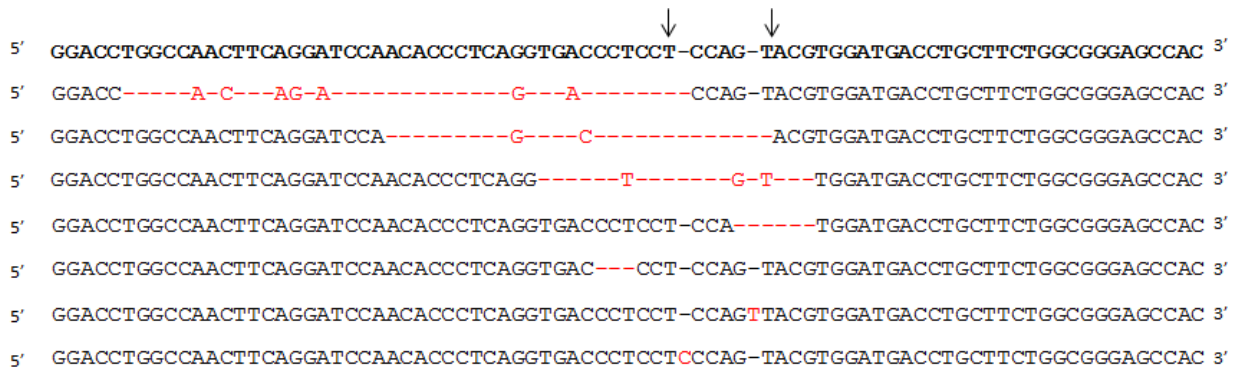


Figure S7: Sanger sequencing validation of the PERV targeting efficiency and indel patterning

We validated the deep sequencing results by Sanger sequencing. Top panel: WT pol sequence. Among 50 clones we checked using Sanger sequencing, 21 of them have deletions/insertions of various forms (bottom panel). The NHEJ pattern and efficiency of 40.5% are consistent with the deep sequencing data.

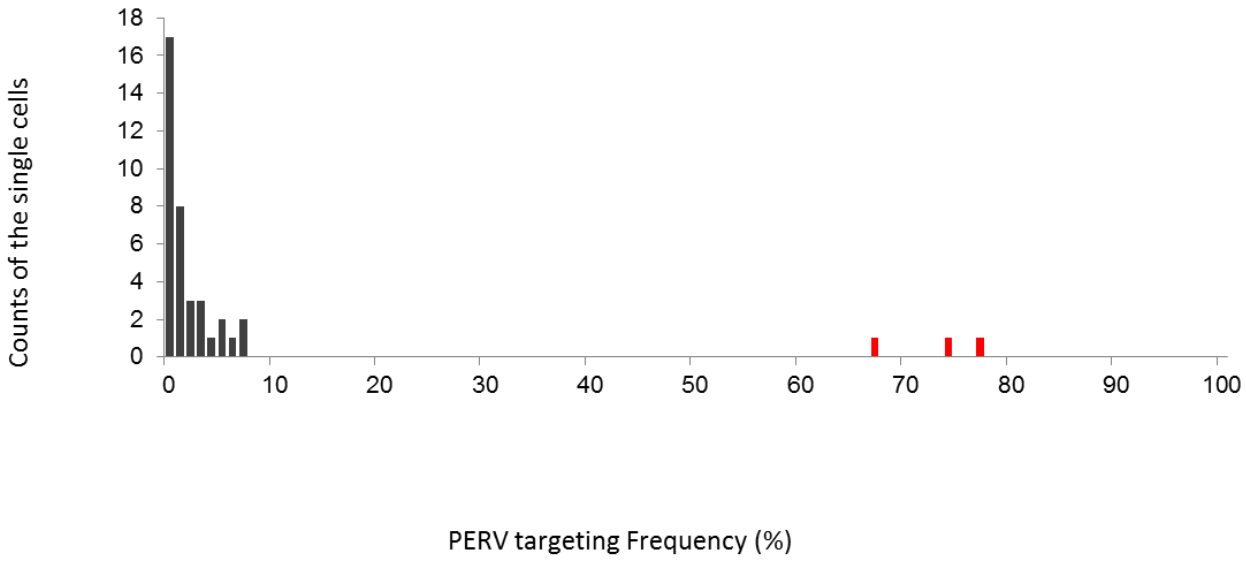


Figure S8: We repeated the gene editing experiment by reestablishing PK15-piggybac-Cas9/2gRNAs, inducing gene editing with DOX for 2 weeks, isolating 40 single cell clones and performing genotyping on the PERV *pol* locus. We observed a reproducible binomial distribution of the PERV editing efficiencies. Among the 40 clones, ~8% (3/40) exhibited 60%-80% PERV targeting efficiency, while the rest of the clones showed < 10% PERV targeting efficiency.

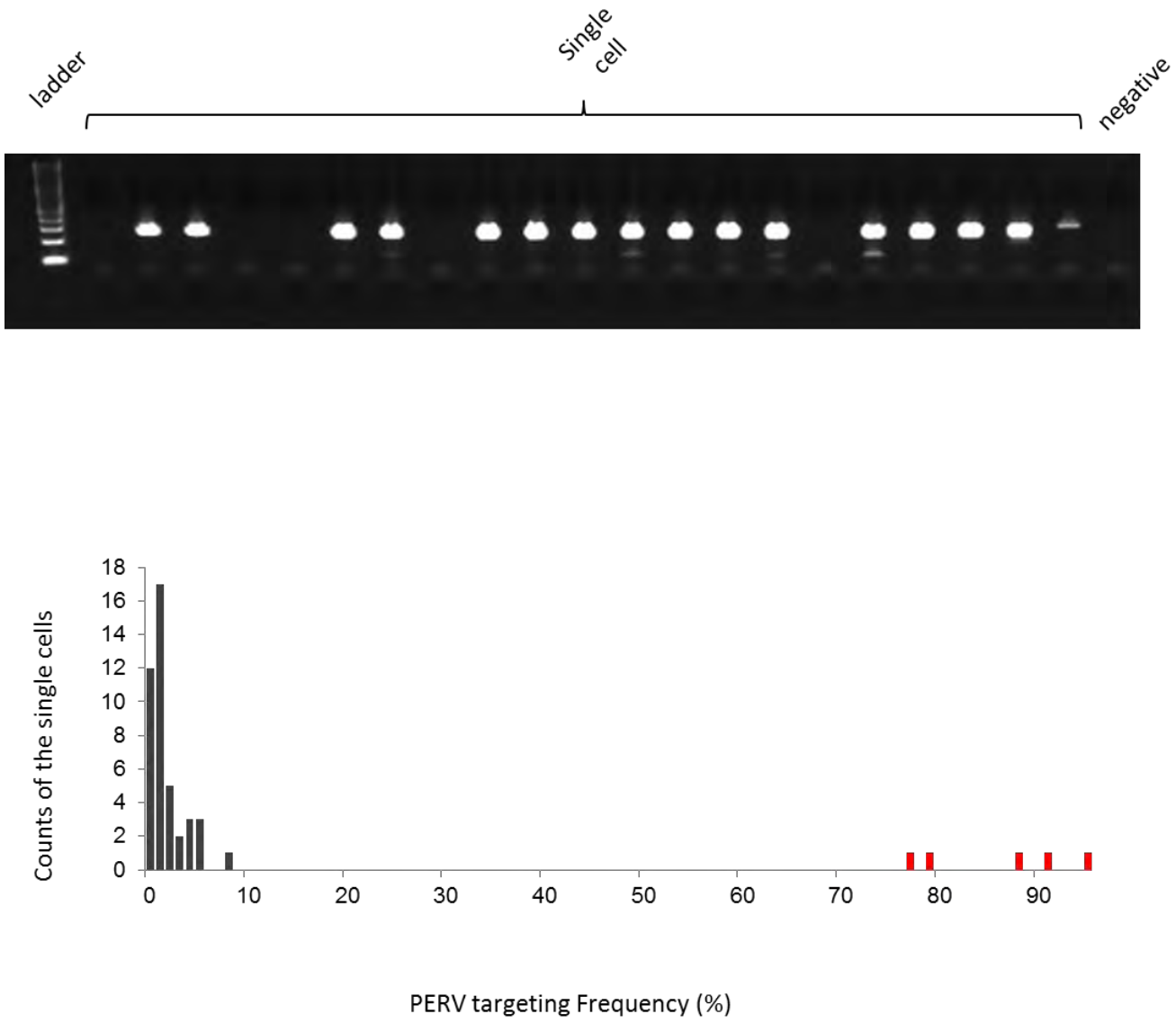


Figure S9: PERV targeting efficiencies of single cells

Genomic DNA was isolated from single cells of PK15 Cas9/2gRNAs after editing and amplified with PERV-specific primers. We attempted to amplify genome from 60 single cell lysis; achieved, on average >80% PCR amplification efficiency of single cell genomic lysis (top) and we observed bimodal distribution of PERV pol editing efficiency among 48 single cells from parents population (bottom).

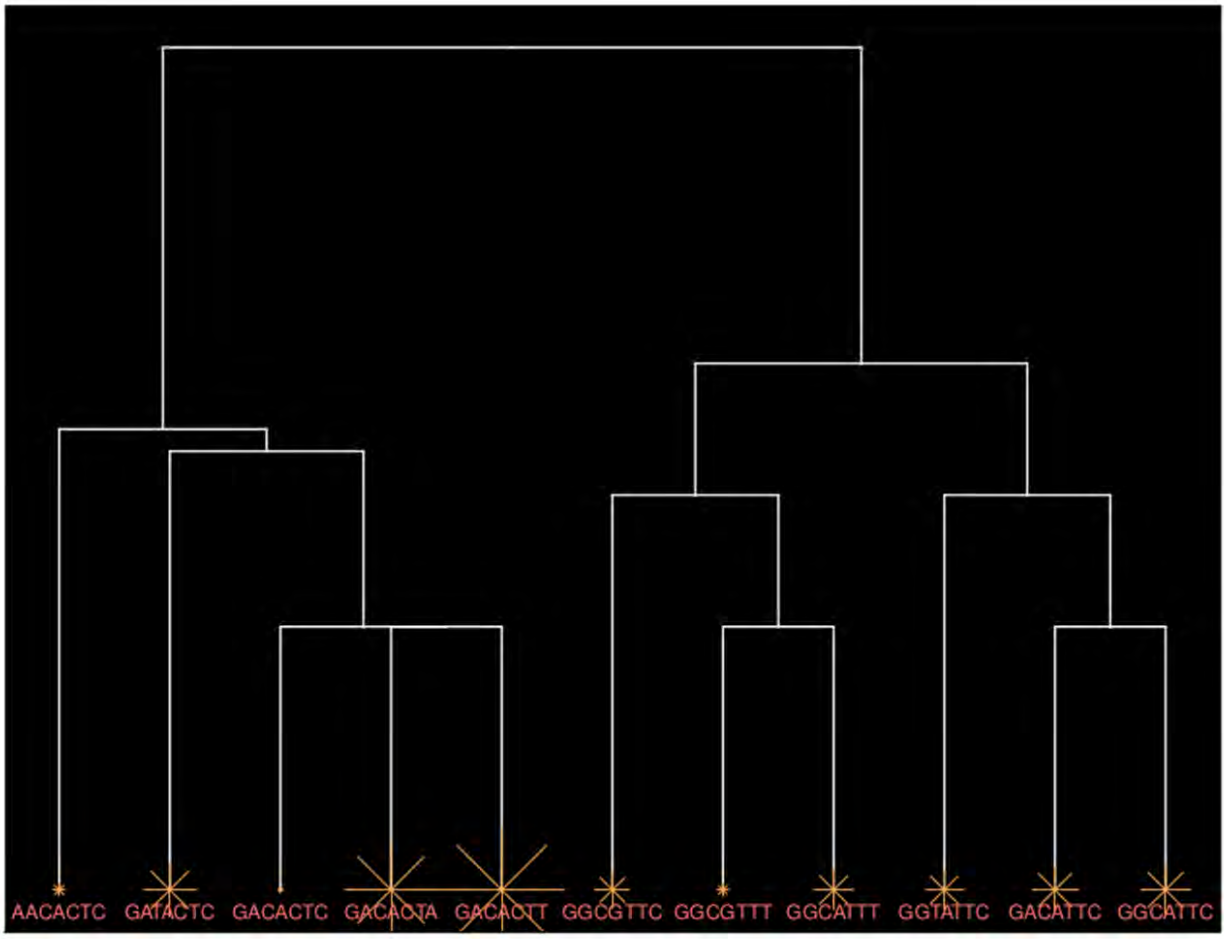


Figure S10: Phylogeny of PERV haplotypes

Viral genomes with different haplotypes have been detected. We observed 7 single nucleotide polymorphisms (SNPs) in the region surrounding the *pol* gene (238bp surrounding the catalytic center). Here, we present the phylogeny of the different haplotypes, and their relative abundance (the yellow star size is proportional to the relative abundance). SNP variants are presented in red.

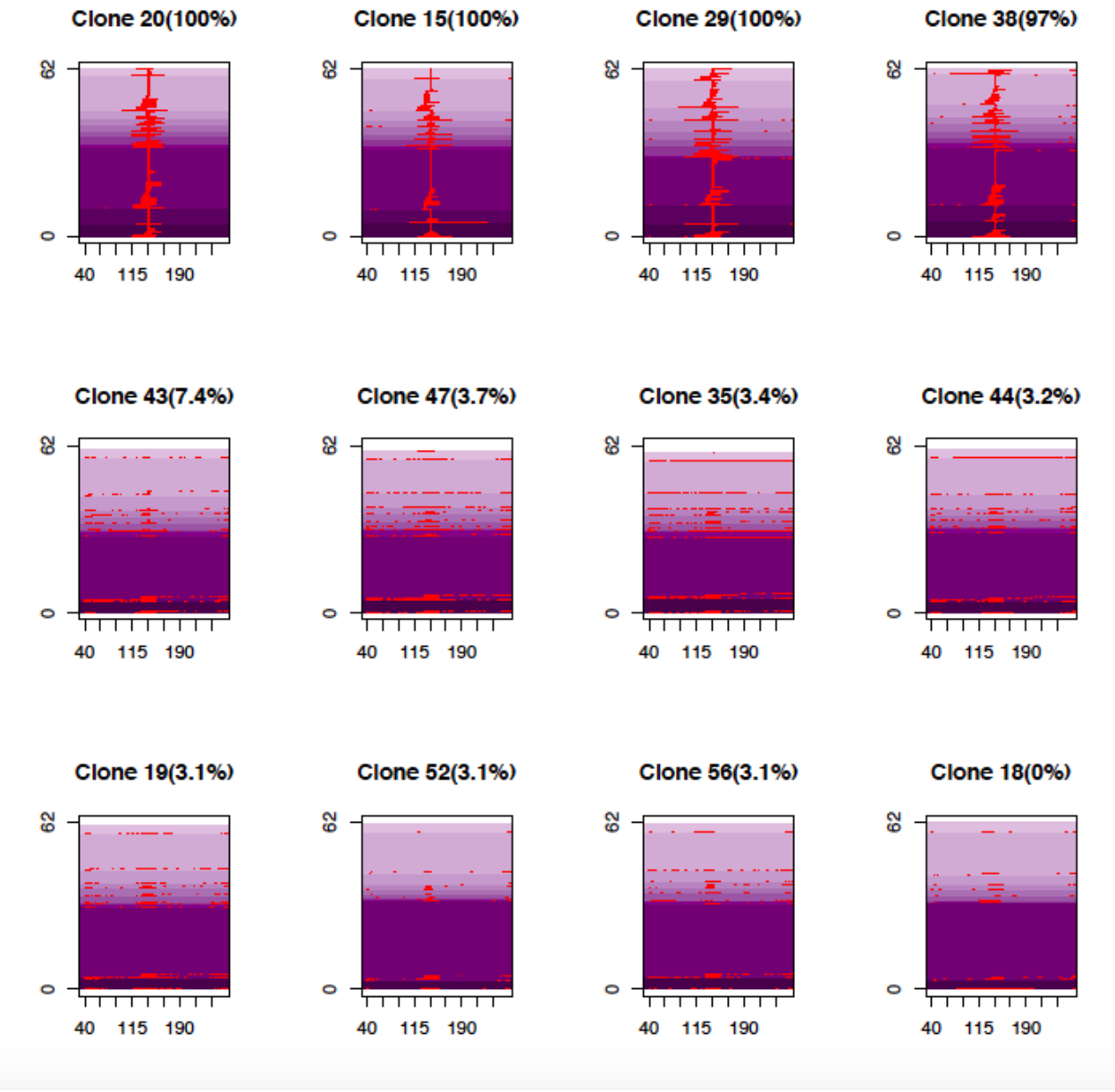
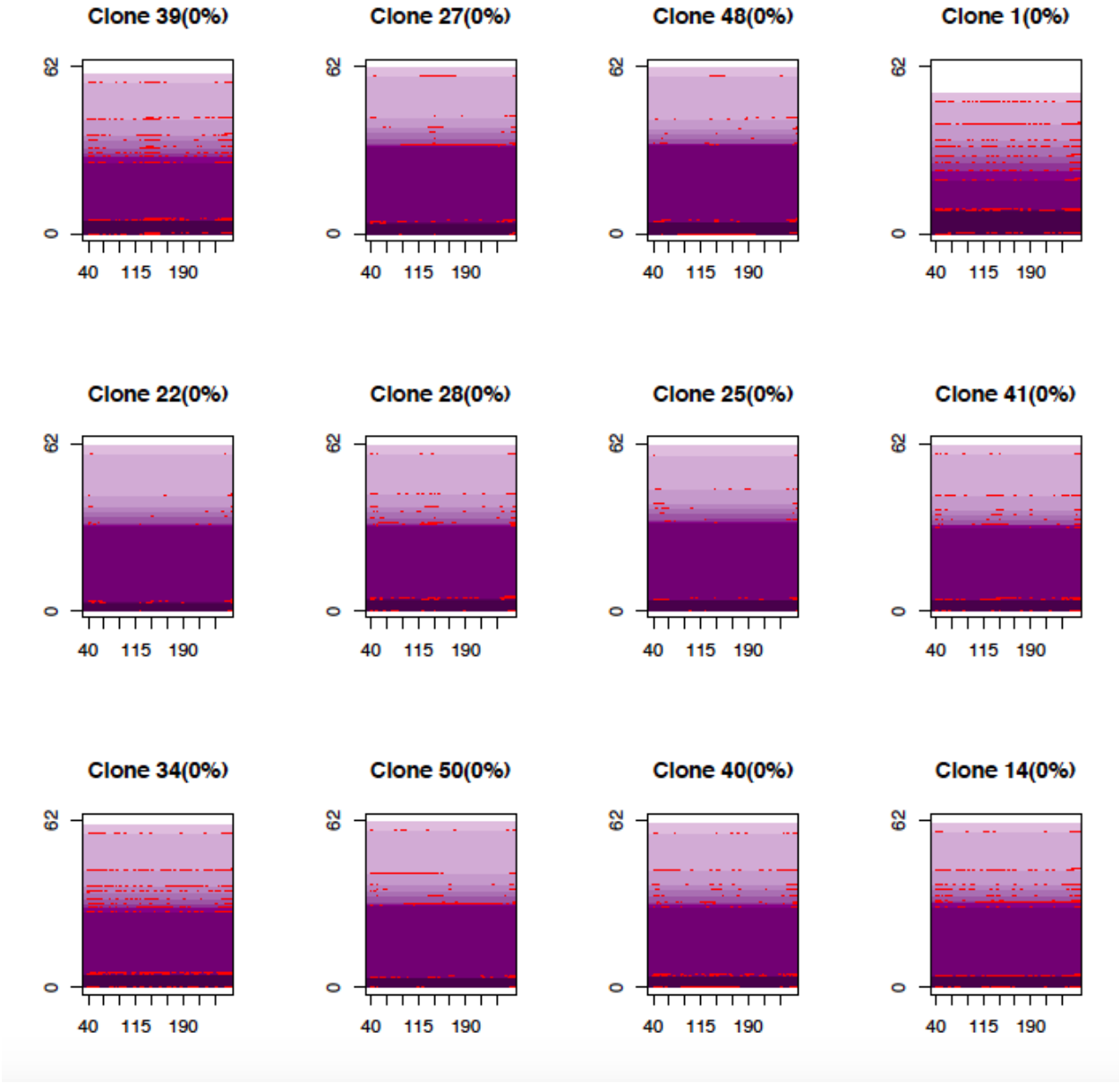
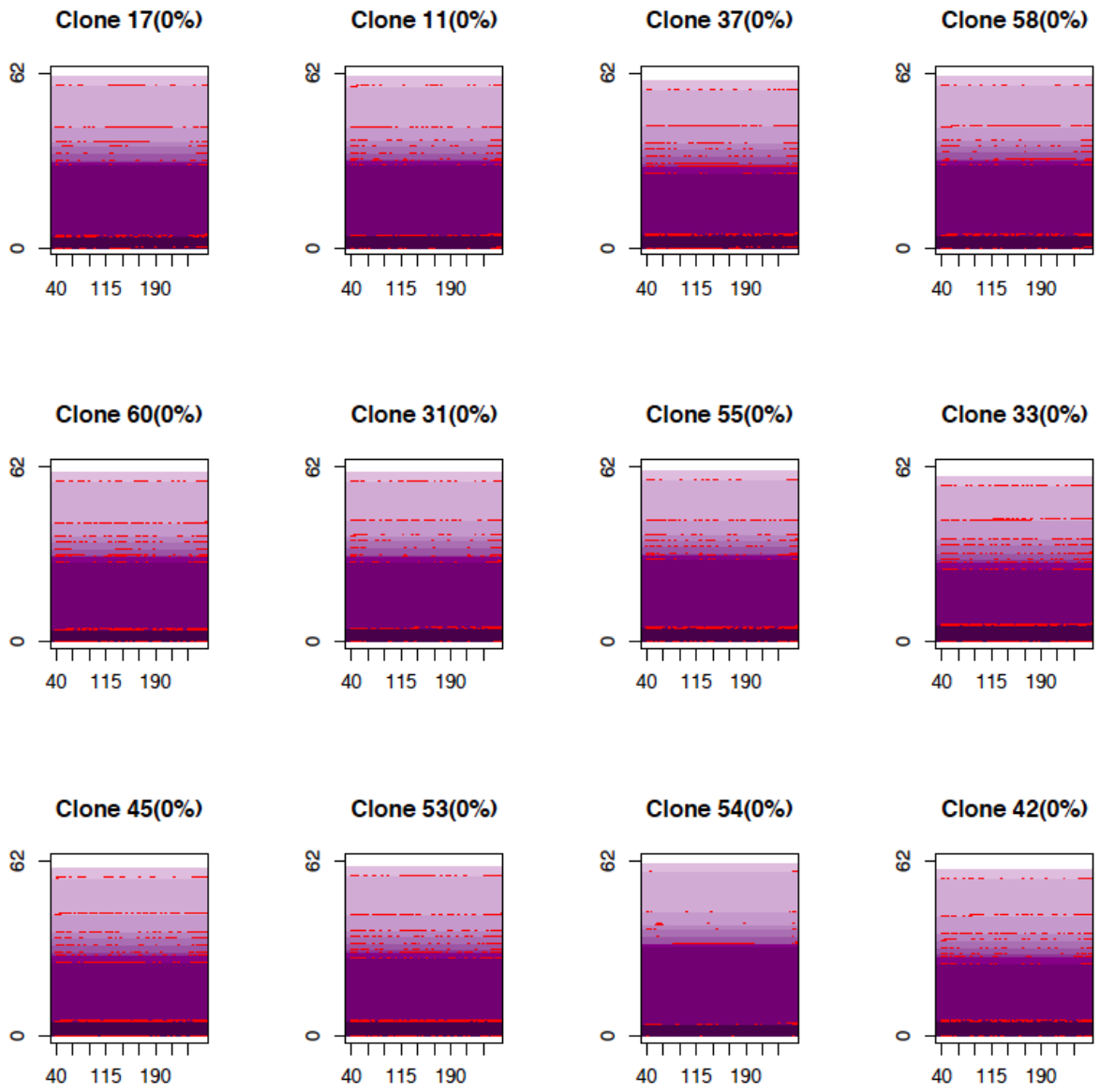


Figure S11: Distribution of the *pol* gene disruption.

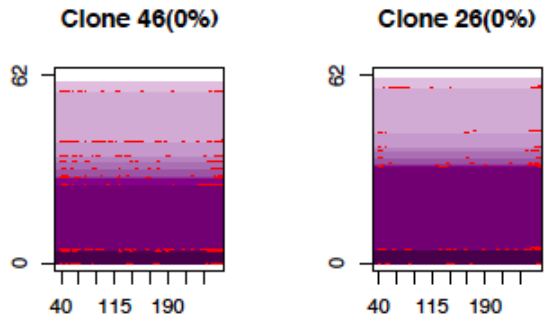
Each plot represents 62 PERV sites for one clone. The y-axis indicates the sites. The X-axis indicates the relative location of the indels within the PERV loci. In red, the aligned indels are shown. Different shades of purple indicate different PERV haplotypes (Figure S7).



Continue: Figure S11



Continue: Figure S11



Continue: Figure S11

a

b

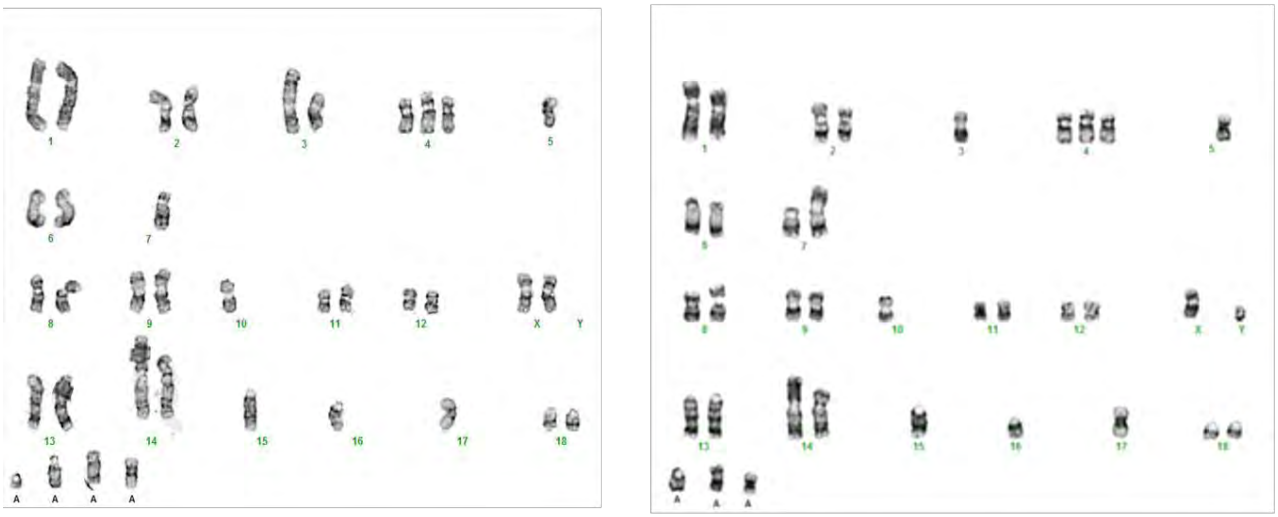


Figure S12: Karyotype analysis of the of the highly and lowly modified PK15 clones

The chromosomal arrangement of one highly modified PK15 clone (a) and one lowly modified PK15 clone (b) were determined using karyotype analysis. “A” represents abnormality (Figure S10-11)

	PK15-1974	CLONE 15	CLONE 20	CLONE 29	CLONE 38	CLONE 40	CLONE 41
Chromosome 1							
Chromosome 2		add(2)(p14.2)		add(2)(p14.2)	add(2)(p14.2)		add(2)(p14.2)
Chromosome 3	add(3)p15	add(3)(p15)	add(3)(p15)	add(3)(p15)	add(3)(p15)	add(3)p15	add(3)(p15)
Chromosome 4		(+4)	(+4,+4)	add(4)(p13.1)	add(4)(p15.1), add(4)(p13.1)	(+4,+4,+4)	(+4)
Chromosome 5	-5	-5	-5	-5	-5	(-5,-5)	-5
Chromosome 6	-6						
Chromosome 7	-7	-7	der(7:16)(q10:q10)	-7	-7	der(7:16)(q10:q10)	-7
Chromosome 8							
Chromosome 9	-9						
Chromosome 10	-10	-10	-10	-10	-10	-10	-10
Chromosome 11	-11						
Chromosome 12							
Chromosome 13							
Chromosome 14	t	der(14;15)(q10;q10), add(14)(p10)	der(14;15)(q10;q10), add(14)(p10)	der(14;15)(q10;q10), add(14)(p10)	der(14;15)(q10;q10), add(14)(p10)	der(14;15)(q10;q10), add(14)(p10)	der(14;15)(q10;q10), add(14)(p10)
Chromosome 15							
Chromosome 16		-16		-16	-16	-16	-16
Chromosome 17	(-17)i(17)(q10)	(-17)i(17)(q10)	(-17)i(17)(q10)	(-17)i(17)(q10)	(-17)i(17)(q10)	(-17)i(17)(q10)	(-17)i(17)(q10)
Chromosome 18							
X,Y	XX	XX	XX	XX	XX	XX	XX
A	n/a	+4mar[8}	+1~5mar[cp10]	+4mar[9]	+3~4mar[cp10]	+1~5mar[cp10]	+4mar[5]

Figure S13: Summary of the karyotype analysis of PK15 highly modified clones (15, 20, 29, 38) and lowly modified clones (40 and 41)

We compared our results with the karyotyping results of the PK15 cell line reported in 1975 (*2*). Most of the chromosomal abnormalities are either from the PK15 strains presented in 1974 report or/all shared by all the strains. The nomenclature can be found in the Figure S11.

Nomenclature	Meaning
+	gain of a chromosome
-	loss of a chromosome
(:)	break (in detailed descriptions)
()	Surround structurally altered chromosomes and breakpoints
add	additional material of unknown origin
der	derivative chromosome
i	isochromosome
p	short arm of a chromosome
q	long arm of a chromosome
t	translocation
a	abnormality

Figure S14: Karyotype nomenclature

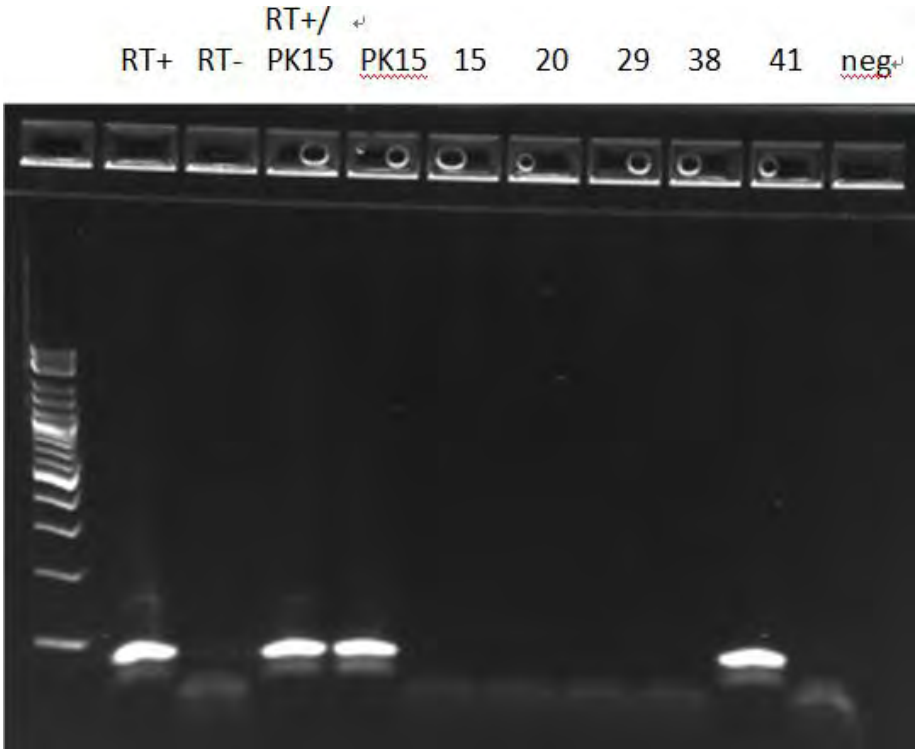


Figure S15: Detection of PERV reverse transcriptase activity in highly and lowly modified PK15 clones and PK15 cells

The sample order from left to right: 2-log DNA ladder (New England Biolabs); RT+ (using commercial reverse transcriptase (RT)); RT- (no RT enzyme); RT+/PK15 (commercial RT enzyme plus PK15 lysis (lysis of virus pellet from PK15 media); PK15 (only PK15 lysis); 15 (only clone 15 lysis); 20 (only clone 20 lysis), 29 (only clone 29 lysis), 38 (only clone 38 lysis), 41(only clone 41 lysis), neg (no lysis or RT enzyme, no RNA template). Clone 15, 20, 29 and 38 are highly modified clones, and 41 is a lowly modified clone.

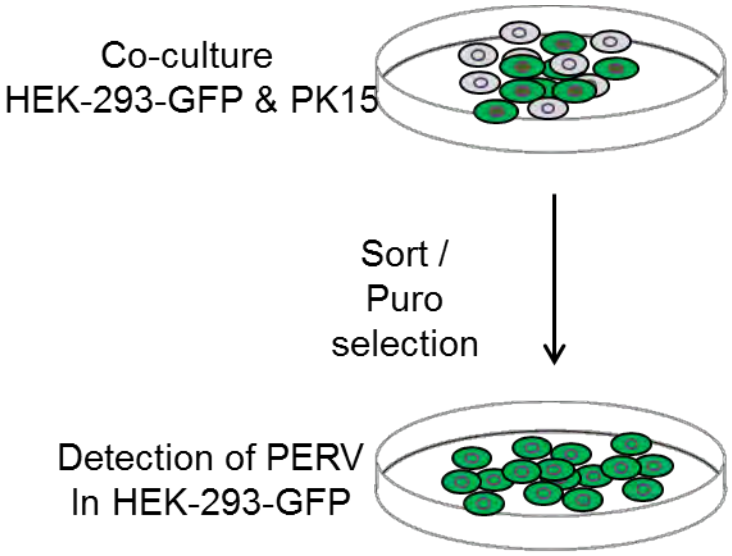


Figure S16: Experimental design to detect the transmission of PERVs to human cells.

HEK-293-GFP cells were co-cultured with equivalent numbers of PK15 cells, the human cell population was isolated by sequential rounds of flow cytometry (sorting based on GFP expression) or by puromycin selection. Genomic DNA of purified human cells was harvested and amplified via PCR to detect and quantify PERV elements in the human cells.

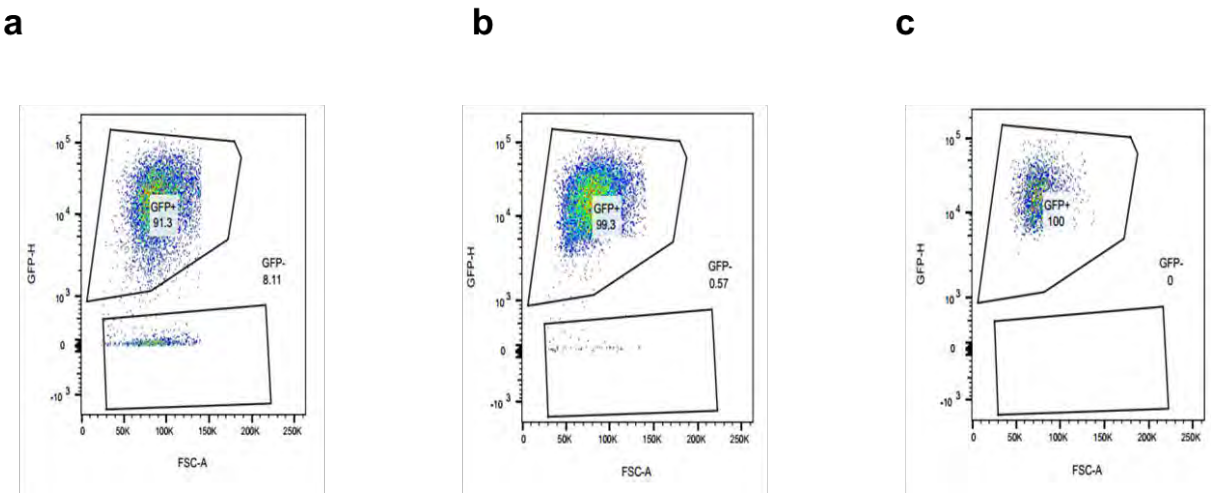


Figure S17: Quality control of the purified HEK293-GFP cells

After co-culturing HEK293-GFP cells with different PK15 clones, we recovered GFP+ cells using several rounds of FACS sorting. We isolated 91.3% GFP cells after the first round of sorting (a), 99.3% after the second round (b), and 100% after the third round (c).

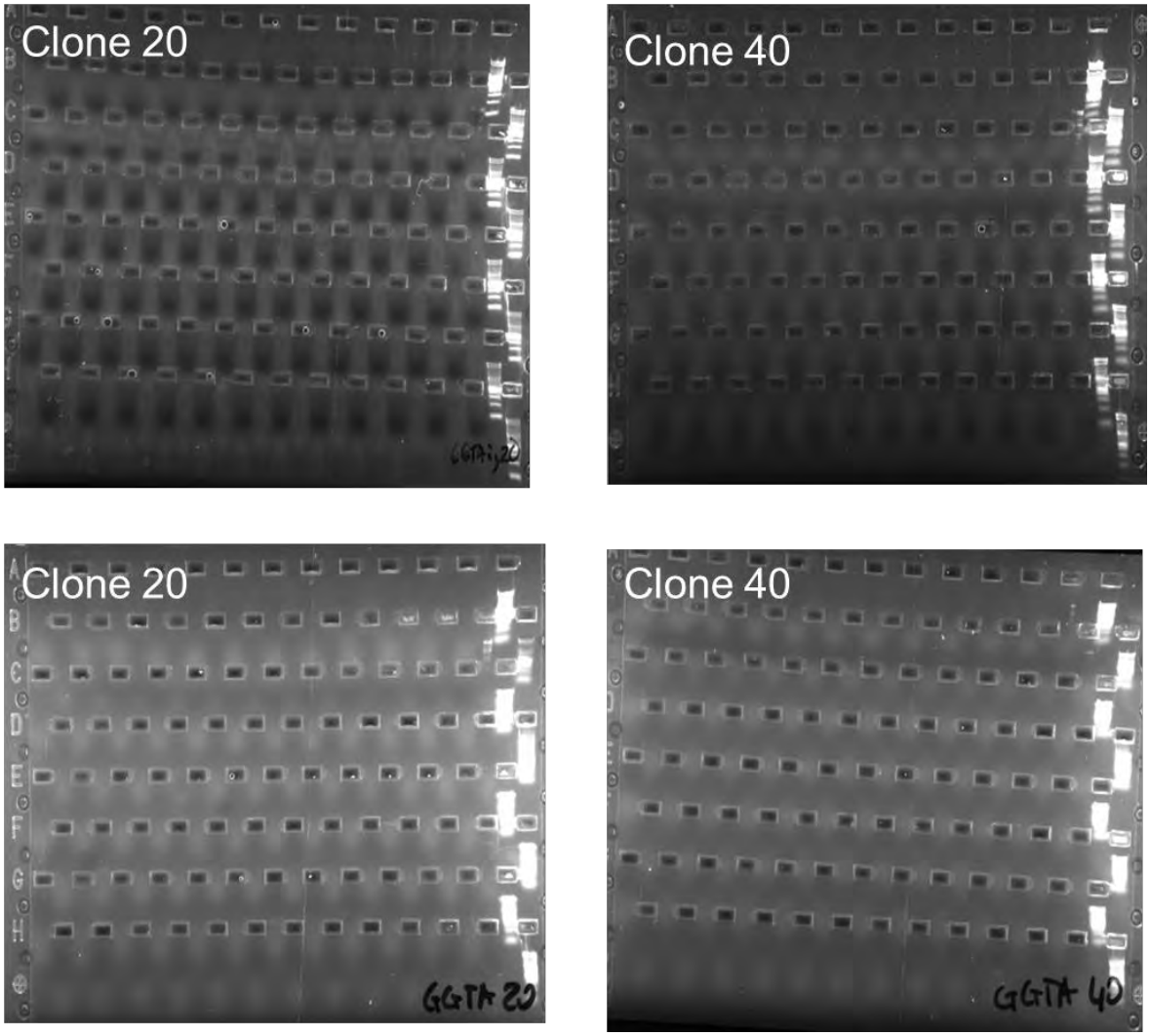


Figure S18: Detection of pig cell contamination in HEK293 cells using pig GGTA1 primers

The top two rows (A and B) contain genomic DNA input equivalent to that of 0.1 cell; C and D equivalent to that of 1 cell, E and F equivalent to that of 10 cells, G and F equivalent to that of 100 cells. The last column in each row is the 2-log DNA ladder (New England Biolabs); the negative control (no DNA template) is shown in second column from the right.

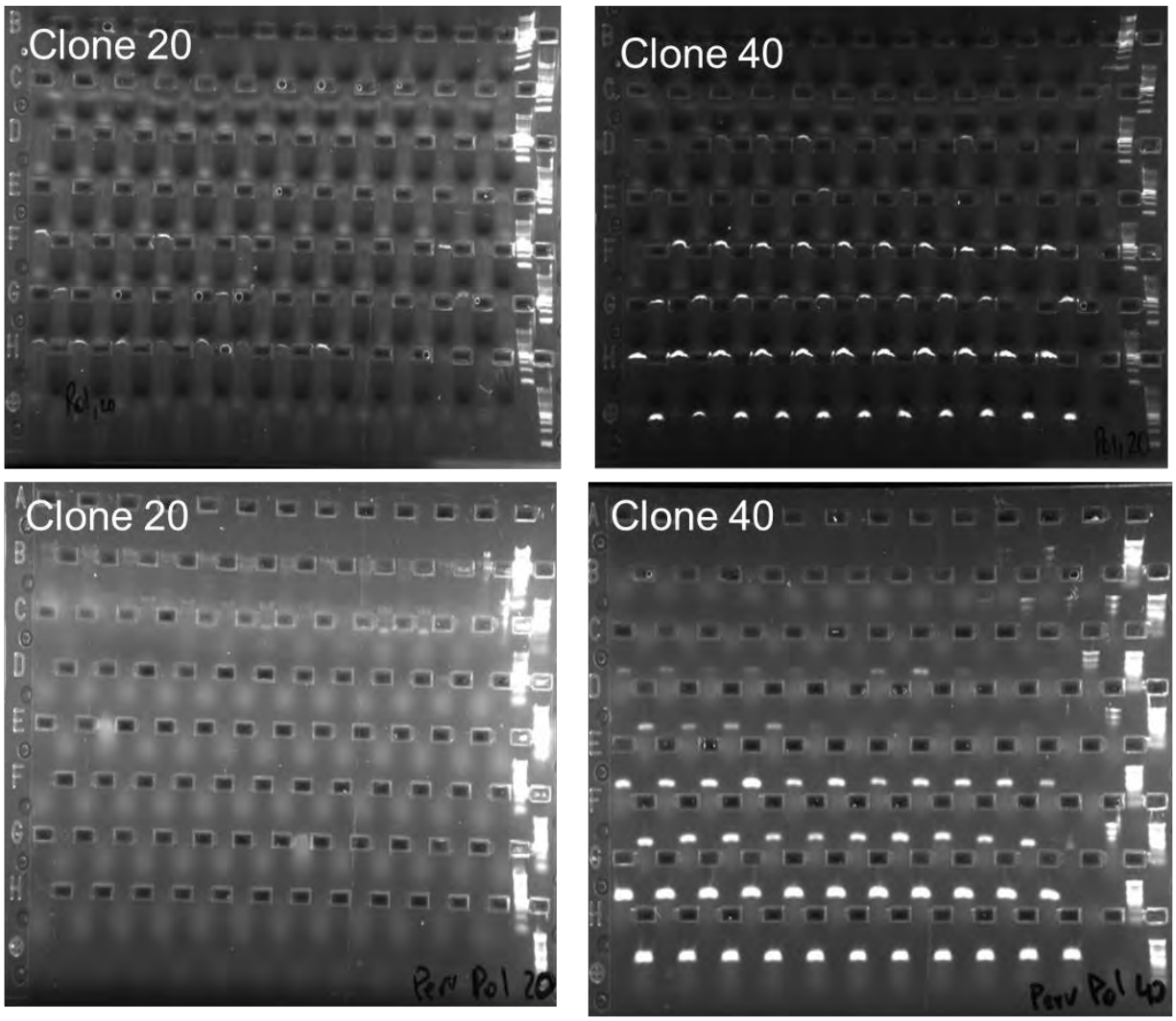


Figure S19: Detection of PERV DNA elements in HEK293 cells using PERV *pop* primers

The top two rows (A and B) contain genomic DNA input equivalent to that of 0.1 cell; C and D equivalent to that of 1 cell, E and F equivalent to that of 10 cells, G and F equivalent to that of 100 cells. The last column in each row is the 2-log DNA ladder (New England Biolabs); the negative control (no DNA template) is shown in second column from the right.

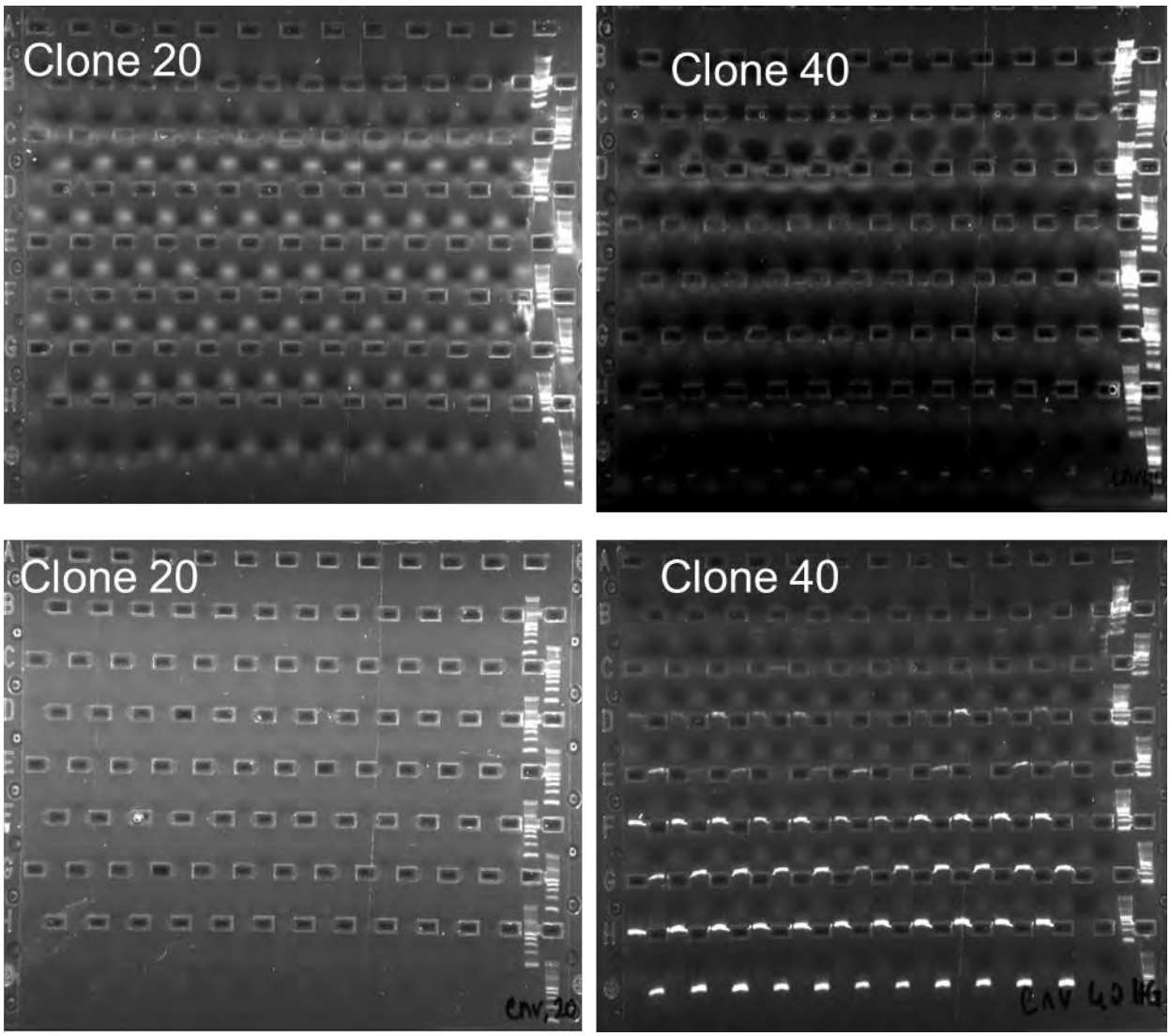


Figure S20: Detection of PERV DNA elements in HEK293 cells using PERV *env* primers

The top two rows (A and B) contains genomic DNA input equivalent to that of 0.1 cells; C and D equivalent to that of 1 cell, E and F equivalent to that of 10 cells, G and F equivalent to that of 100 cells. The last column in each row is the 2-log DNA ladder (New England Biolabs); the negative control (no DNA template) is shown in second column from the right.

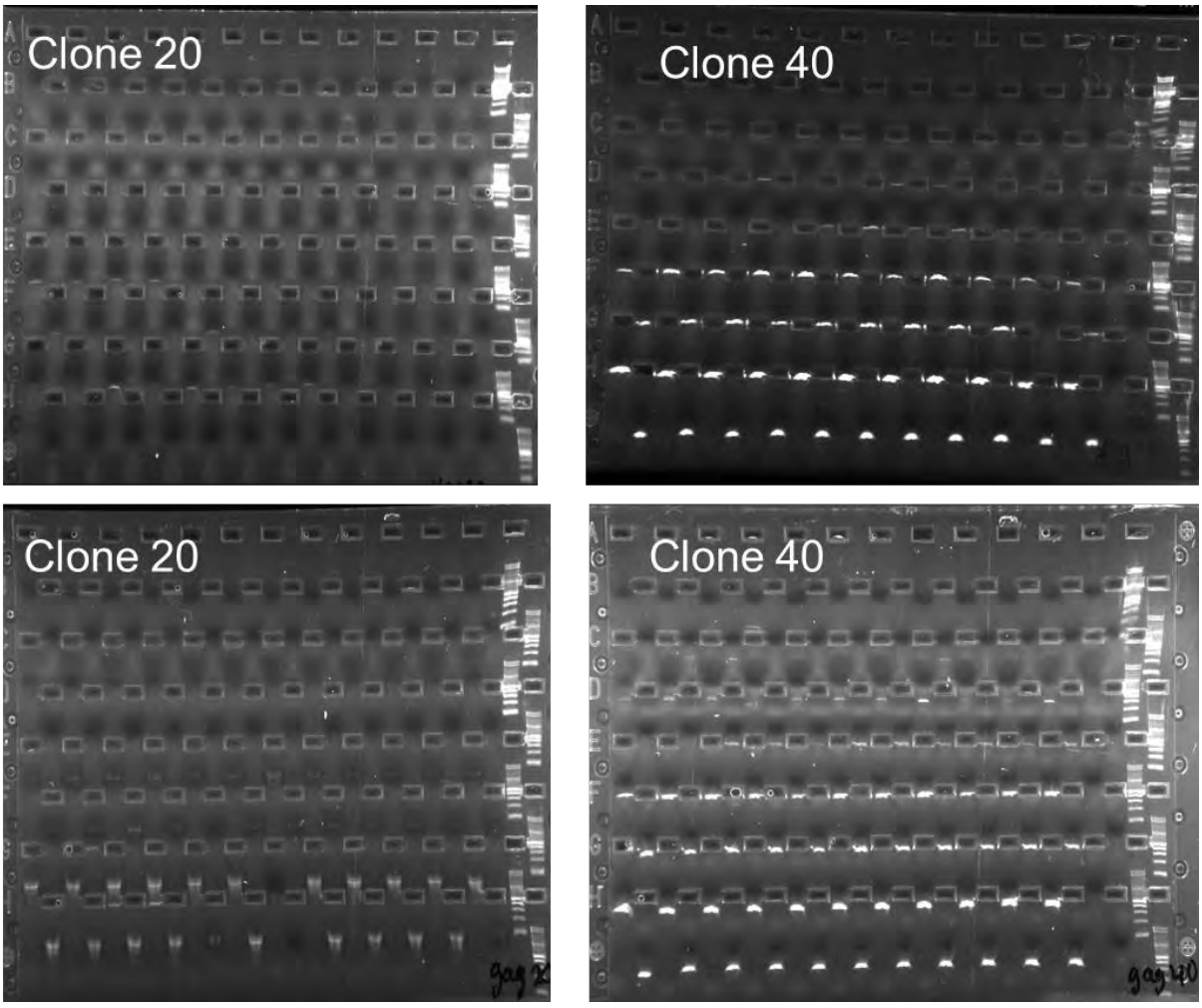


Figure S21: Detection of PERV DNA elements in HEK293 cells using PERV *gag* primers

The top two rows (A and B) contains genomic DNA input equivalent to that of 0.1 cell; C and D equivalent to that of 1 cell, E and F equivalent to that of 10 cells, G and F equivalent to that of 100 cells. The last column in each row is the 2-log DNA ladder (New England Biolabs); the negative control (no DNA template) is shown in second column from the right.

a

	Clone 15	Clone 20	Clone 29	Clone 38	Clone 40	Clone 41	PK15
Cas9	782	771	207	416	61	78	0
gRNA1	5	3	2	1	0	0	0
gRNA2	4	4	2	1	0	0	0
Total	75815790	72609563	82295929	72569325	60990510	80508090	0

b

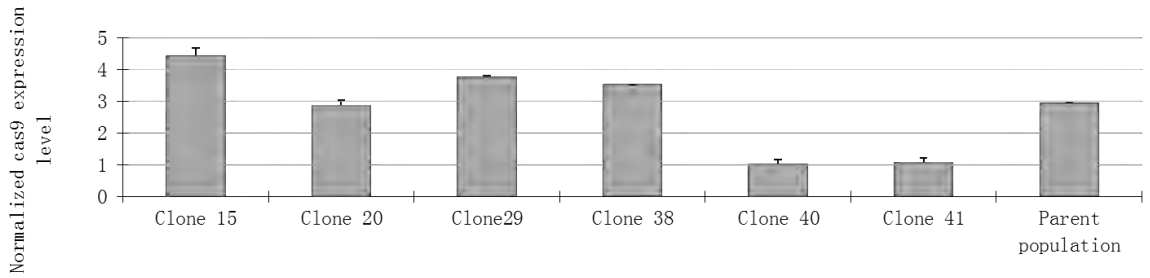


Figure S22: Cas9/2gRNAs expression levels in highly and lowly modified clones

a, Transcriptome data indicated that Cas9-gRNA expression level is higher in highly modified clones (15, 20, 29, 38) compared with lowly modified clones (40, 41). The table shows the counts of reads in the RNA-seq data. We observed a higher cas9 expression, and gRNAs in highly modified clones using a t-test of the normalized and scaled gene expression (*P-value(cas9)=0.022, P-value(gRNA1)=0.024,P-value(gRNA2)=0.017, one-tail*). b. We validated the RNA-seq data with RT-qPCR to examine the Cas9 expression level among the clones. We observed significantly higher Cas9 expression in PERV pol highly modified clones than lowly modified clones (ttest, p-value= 0.03, one-tail).

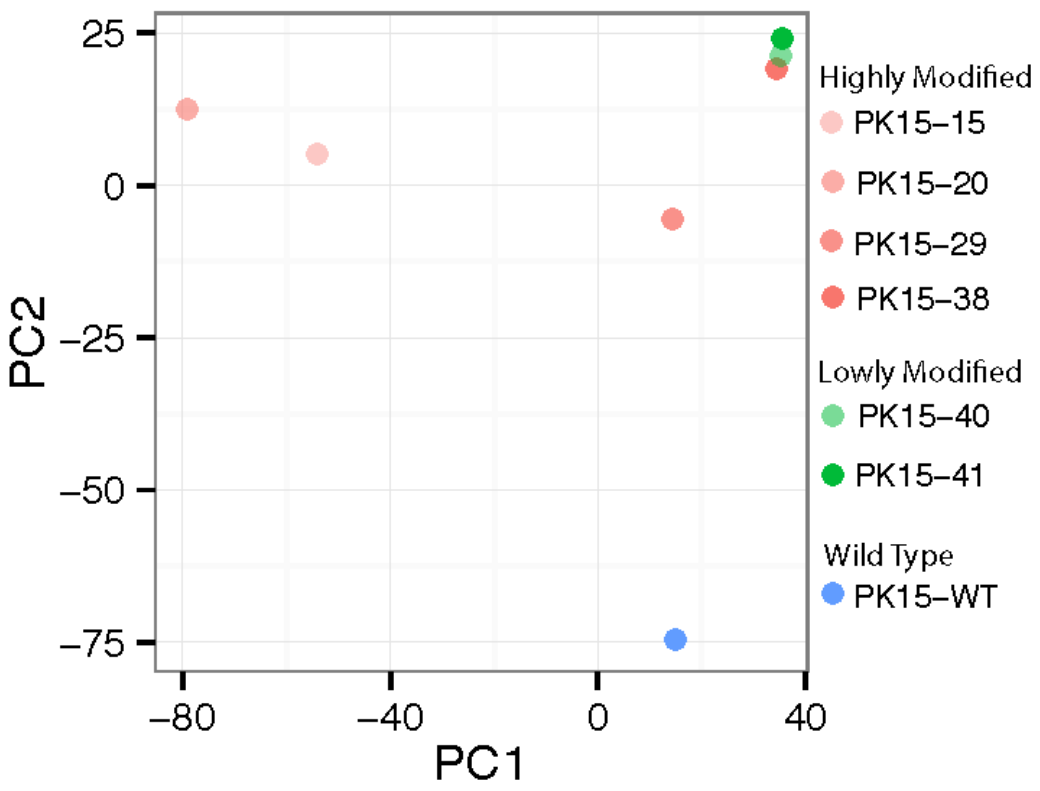


Figure S23: Principle component analysis of highly and lowly modified PK15 clones
Principle component analysis revealed that highly modified clones (red) have much more diversified gene expression profiles than lowly modified clones do (green).

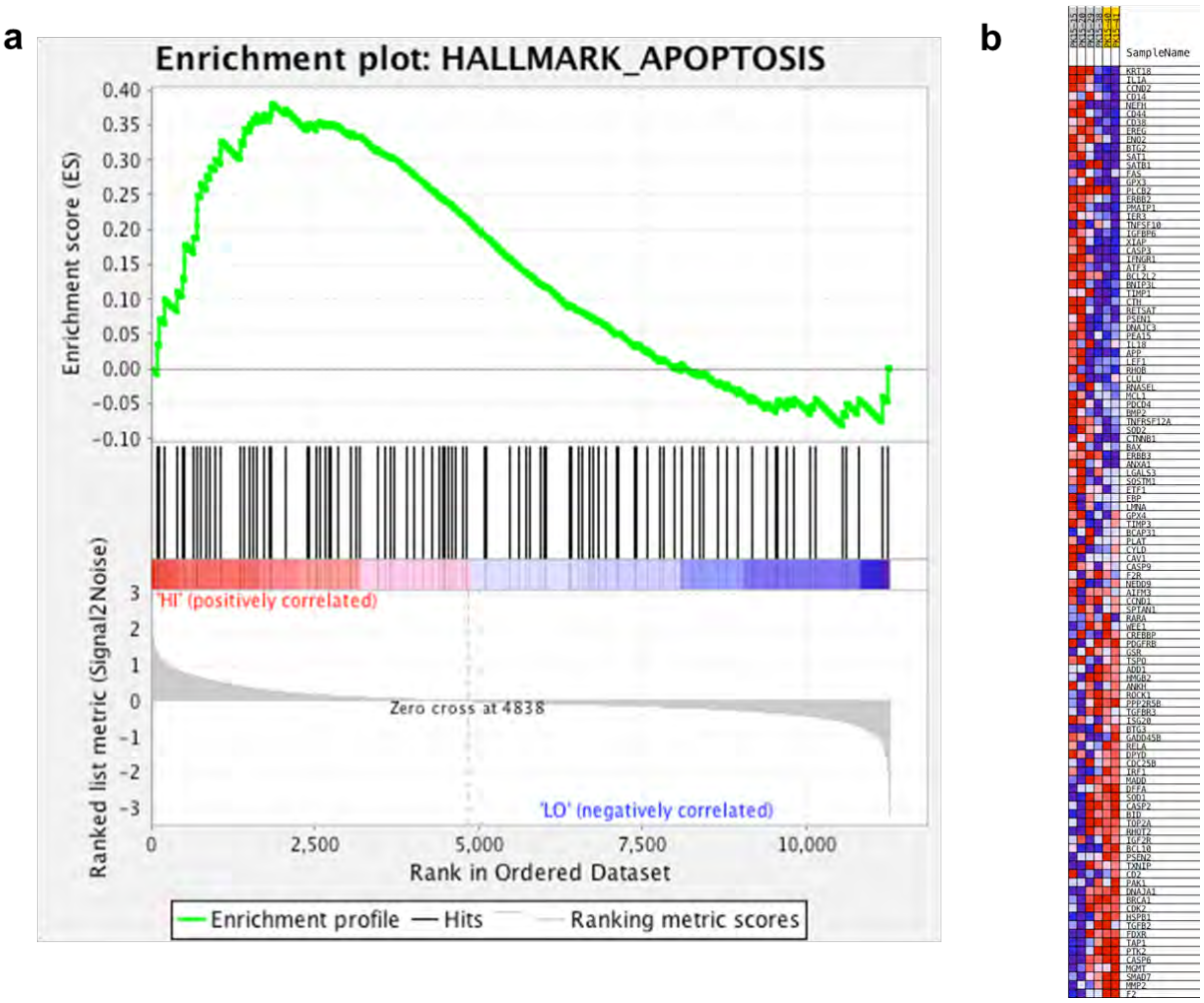


Figure S24: Gene set enrichment analysis

Gene set enrichment analysis revealed that the hallmark apoptosis gene set was significantly up-regulated in high modified clones compared with low modified ones (nominal *P-value=0.03* and *Q-value=0.08*). A) Enrichment score (ES) and rank order of the genes in the hallmark apoptosis set. B) Heat map of the genes in the hallmark apoptosis set.

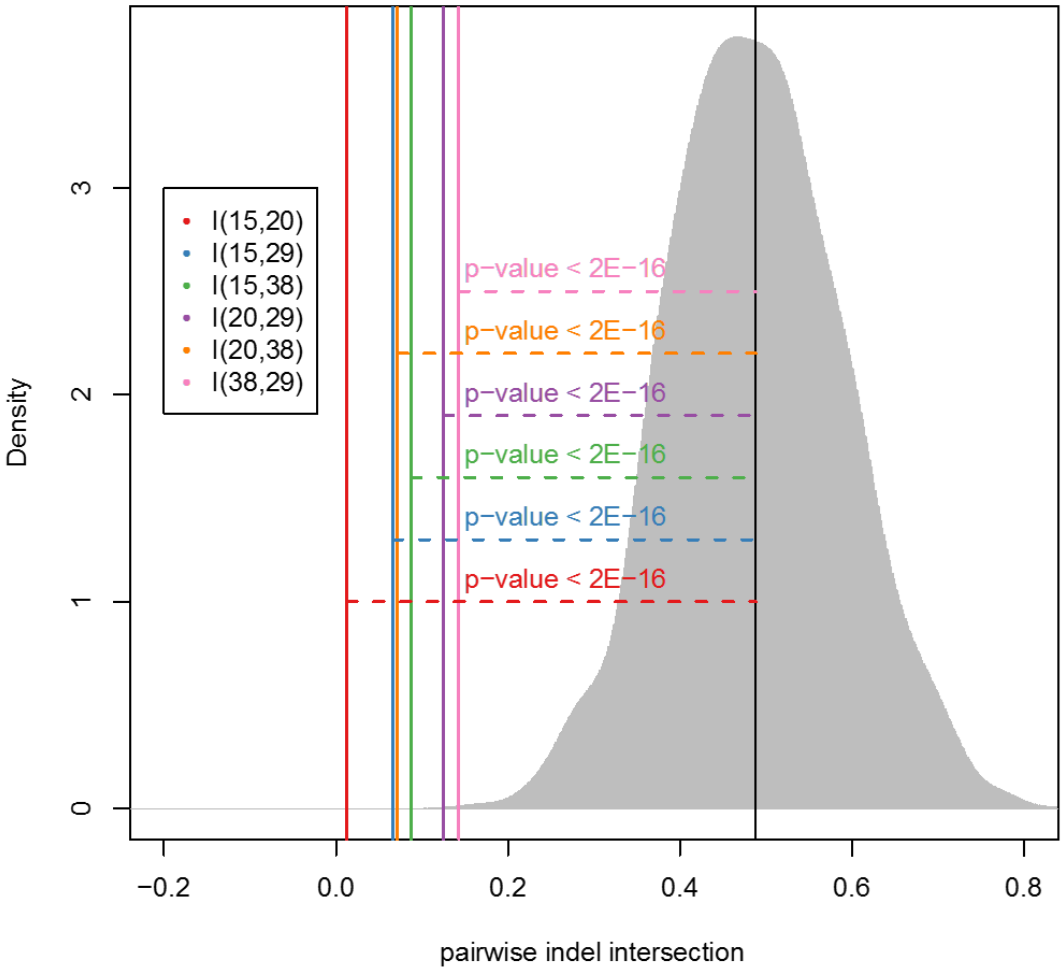


Fig. S25. Indels composition analysis and comparison between highly modified clones.

We measured the pairwise level of intersection of indels between the 4 highly edited clones (I(15,20), I(15,38), I(15,29), I(20,38), I(20,29), I(20,38); colored vertical lines). We compared these measured levels of intersection to a null distribution (grey). The null distribution was constructed by repeating 1000 times the estimation pairwise of indels between the 4 highly edited clones with the indels clone labels randomized. Level of intersection is defined as the ration of distinct indels that are present in two given clones respect to the total number of indels. We carried out one-sided t-tests (alternative="greater") to compare the null distribution to the different levels of intersection for each of the clone pairs.

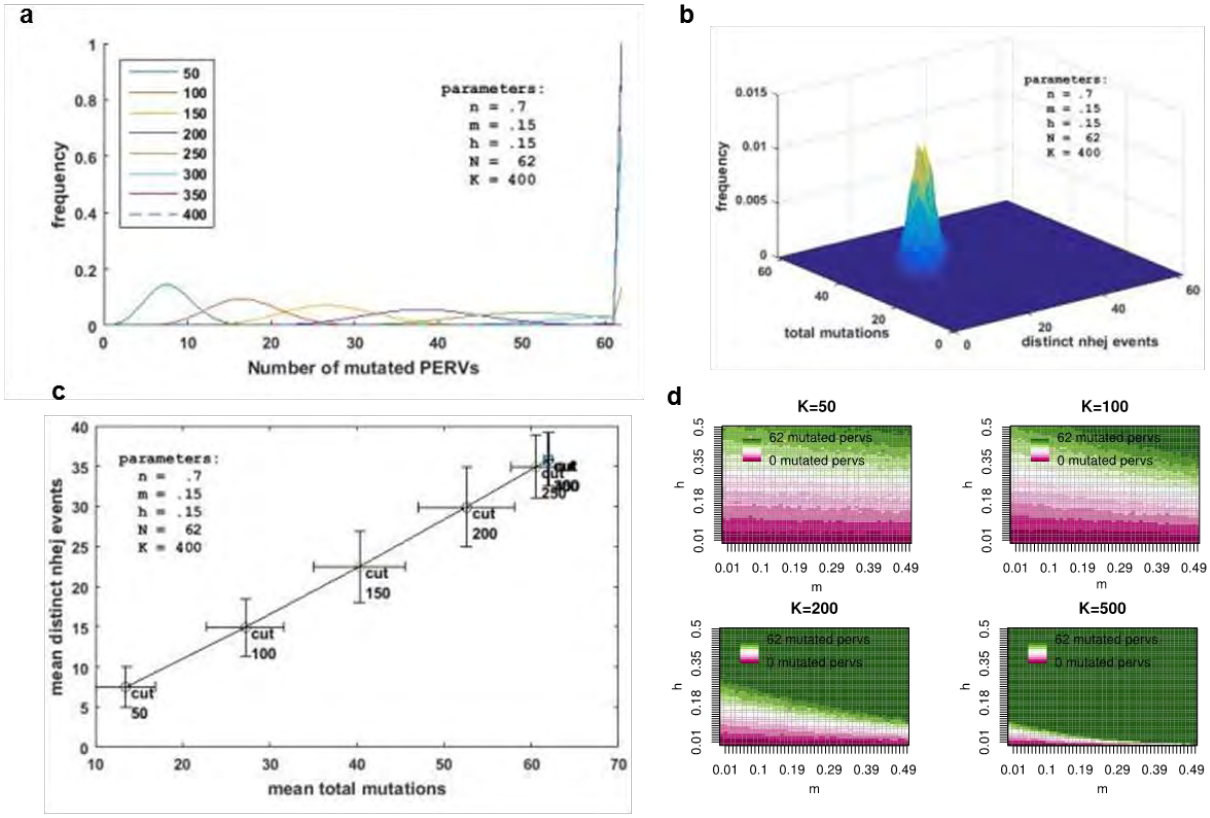


Figure S26: Markov model analysis of DNA repair processes leading to Cas9 elimination of active PERV elements.

We analyzed the interaction between error-free NHEJ repair (with probability *n*), mutagenic NHEJ repair (with probability *m*), and HR repair (with probability *h*) of Cas9-cut PERV elements using a Markov model (see Methods). The model computes the distribution of the number of mutated PERV elements after a given number of Cas9 cuts. a-c) For all values of *n*, *m*, and *h* tested (see Methods), this distribution is unimodal and has a mean that increases monotonically with the number of cuts, and that ultimately leads to elimination of all PERV elements when the number of cuts becomes sufficiently large. a) provides an illustration of such an advancing unimodal mutation distribution for the indicated values of *n*, *m*, and *h* that proceeds to PERV elimination over 400 cuts. (Continue)

Continue Fig. S26: Markov model analysis of DNA repair processes leading to Cas9 elimination of active PERV elements.

b) shows the bivariate distribution of total mutations *vs.* distinct NHEJ events predicted by the model at cut 150. c) shows that the means of total mutations and distinct NHEJ events increase in a linear fashion with the number of cuts until all PERVs are eliminated (error bars: 1 standard deviation). The unimodality observed in a) suggests that the bimodal distribution of the number of inactivated PERV elements observed in Cas9-edited PK15 clones (see Figure 2) is not due to the stochastics of DNA repair alone, but likely reflects a bimodal state of the initial cells. Two possible hypotheses are: (i) There is a bimodal state regarding repair processes, in which (e.g.) one subset of cells has a very low *n* or high *m* relative to the rest, and (ii) Cas9 expression (and thus, the number of cuts made by Cas9) is high in one fraction and low in the rest. Both our observations of elevated Cas9 and gRNA expression levels in the highly modified clones (Fig. S23), and that the hallmark apoptosis gene set was also significantly up-regulated in these clones (Fig. S25), support hypothesis (ii). Indeed, the upregulation of apoptosis of Fig S23 suggests high stress in highly modified clones that could be directly explained by high levels of Cas9 causing more DNA cut and repair events. By contrast, hypothesis (i) would imply that the bimodal mutation distribution of Fig. 2 arises because, among cells that experience the same number of cuts, some are prone to repairing them *via* perfect *vs.* mutagenic NHEJ, while the rest have the opposite tendency. This would not explain the high levels of Cas9 and apoptotic stress in the highly edited clones. d) Depiction of the most likely number of mutations predicted by the model based on a grid of 2500 sets of *n*, *m*, and *h* values covering their theoretical ranges, shown for four number of cuts (*K*): 50, 100, 200, 500 (see Methods). Notably, at any number of cuts, and for any value of *m*, increases in *h* lead to higher predicted numbers of mutations, suggesting that HR always expedites PERV elimination. The model clarifies that HR operates to help ratchet up the number of inactivating PERV mutations because, as only wild-type targets can be recognized by Cas9, HR can at best restore the cut to wild-type using a wild-type template, but will otherwise mutate it by copying in a previously mutated site.

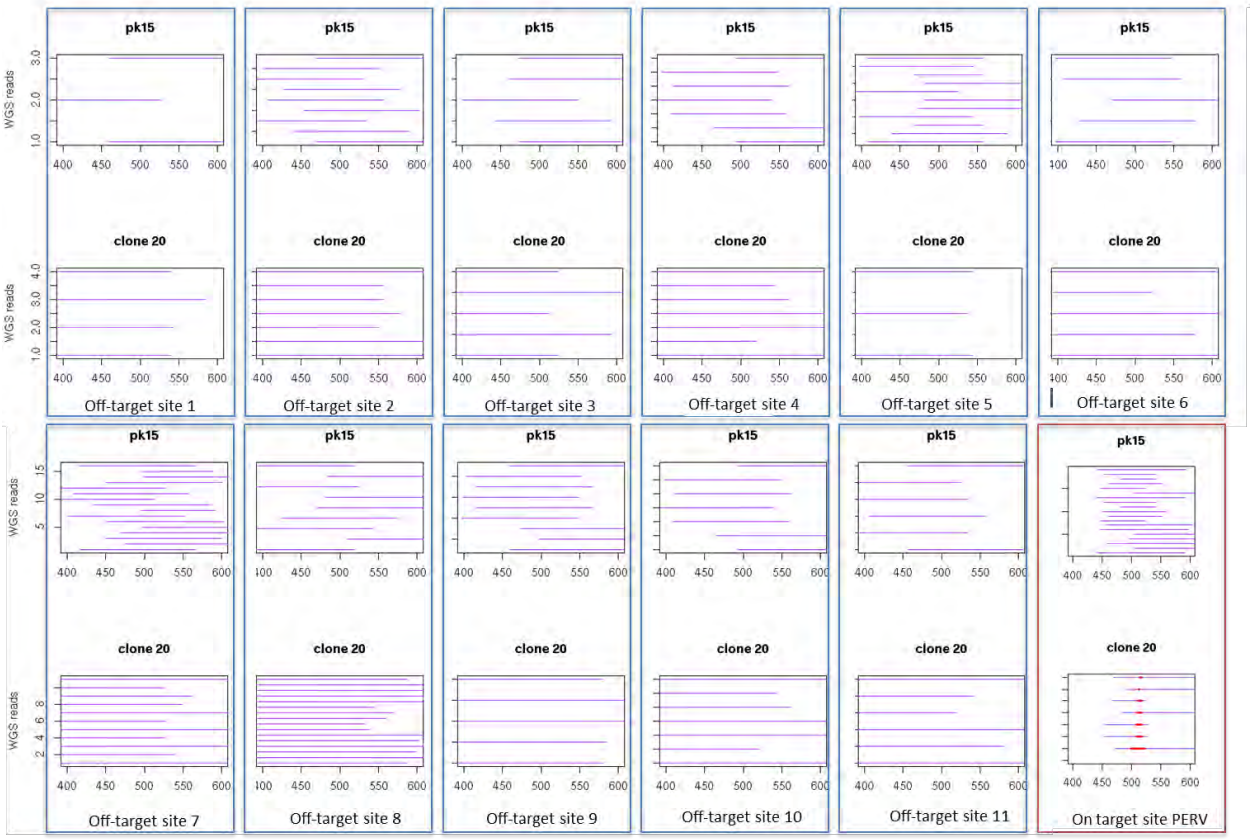


Fig. S27. Off-target analysis using Whole Genome Sequencing (WGS)

We looked for mutations in the *Sus Scrofa* reference sequence (ver 10.2) up to 2 mismatches away from the two 20 bp Cas9 targets. We identified 11 sites up to 2bp away from the target and extracted them together with 500 bp of their neighboring regions. We used BLAT to map the WGS reads to the extracted reference sequences and searched for potential indel patterns that had emerged in Clone 20 as a result of off-target effects. We obtained an average coverage of 7-8X per locus, after excuding reads with <50 bp matches or more than 0 mismatches with the reference sequence, and reads that did not mapped together with its mate read. After inspecting the remaining mapped reads, we did not detect any off-target indel patterns present in clone 20 in the off target regions (blue boxes), whereas we could see very clearly in all the loci with perfect gRNA match (red box). In purple, reads matching the reference are displayed, in red, indels detected.

Methods:

PERV copy number quantification: We used Droplet Digital PCR ™ PCR (ddPCR™) to quantify the copy number of PERVs according to the manufacturer's instructions (Bio-Rad). Briefly, we purified genomic DNA (DNeasy Blood & Tissue Kit, Qiagen) from cultured cells, digested 50 ng genomic DNA with MseI (10U) at 37 ºC for 1 hour, and prepared the ddPCR reaction with 10 µl 2X ddPCR Master mix, 1 µl of 18µM target primers & 5µM target probe (VIC), 1µl of 18µM reference primers & 5µM reference probe (FAM), 5ng digested DNA, and water to total volume of 20 µl. The sequence of the primers and the probe information can be found in Extended Data Table 1.

Methods Table 1- Primers used in ddPCR assay

Name	**Sequence**
PrimerPol1-FW	CGACTGCCCCAAGGGTTCAA
PrimerPol2-FW	CCGACTGCCCCAAGAGTTCAA
PrimerPol-RV	TCTCTCCTGCAAATCTGGGCC
ProbePol	/56FAM/CACGTACTGGAGGAGGGTCACCTG
Primerpig_actin_F	taaccgatcctttcaagcattt
Primerpig_actin_R	tggtttcaaagcttgcatcata
Probepig_actin	/5Hex/cgtggggatgcttcctgagaaag
Primerpig_GAPDH_F	ccgcgatctaatgttctctttc
Primerpig_GAPDH_R	ttcactccgaccttcaccat
Probepig_GAPDH	/5Hex/cagccgcgtccctgagacac

CRISPR-Cas9 gRNAs design: We used MUSCLE (*1*) to carry out a multiple sequence alignment of 245 endogenous retrovirus found in the porcine genome. We built a phylogenetic tree of the sequences and identified a clade that included the PERVs (see Fig. 1a). We used the R library DECIPHER to design specific gRNAs that target all PERVs but no other endogenous retroviral sequences.

Cell culture: PK15 were maintained in Dulbecco's modified Eagle's medium (DMEM,

Invitrogen) high glucose supplemented with 10% fetal bovine serum (Invitrogen), and 1% penicillin/streptomycin (Pen/Strep, Invitrogen). All cells were maintained in a humidified incubator at 37^{o}C and 5% CO_2.

PiggyBac-Cas9/2gRNAs construction and cell line establishment: PiggyBac-Cas9/2gRNAs construct is derived from a plasmid previously reported in Wang et al (*2*). Briefly, we synthesized a DNA fragment encoding U6-gRNA1-U6-gRNA2 (Genewiz) and incorporated it into a PiggBac-Cas9 construct. To establish PK15 cell lines with PiggyBac-Cas9/2gRNAs integration, we transfected $5{\cdot}10^5$ PK15 cells with 4 µg PiggyBac-Cas9/2gRNAs plasmid and 1 µg Super PiggyBac Transposase plasmid (System Biosciences) using Lipofectamine 2000 (Invitrogen). To enrich for the cells carrying the integrated construct, we added 2 µg/mL puromycin to the transfected cells. Based on the negative control, in which we applied puromycin to wild type PK15 cells, we determined that the selection completed in 3 days. The PK15-PiggyBac cell lines were maintained with 2 µg/mL puromycin hereafter. 2 µg/ml doxycycline was applied to induce Cas9 expression.

Lentivirus-Cas9/2gRNAs construction and cell line establishment: Lenti-Cas9/2gRNAs constructs were derived from a plasmid previously reported (*3*). We synthesized a DNA fragment encoding U6-gRNA1-U6-gRNA2 (Genewiz) and incorporated it into a Lenti-Cas9-V2. To generate lentivirus carrying Lenti-Cas9/2gRNAs, we transfected ~ $5{\cdot}10^6$ 293FT HEK cells with 3 µg Lenti-Cas9-gRNAs and 12 µg ViraPower Lentiviral Packaging Mix (Invitrogen) using Lipofectamine 2000. The lentiviral particles were collected 72 hours after transfection, and the viral titer was measured using Lenti-X GoStix (Takara Clonetech). We transduced ~ 10^5 lentiviral particles to ~ $1{\cdot}10^6$ PK15 cells and conducted selection by puromycin to enrich transduced cells 5 days after transduction. The PK15-Lenti cell lines were maintained with 2 µg/mL puromycin thereafter.

Genotyping of colonized and single PK15 cells: PK15 cultures were dissociated using TrypLE (Invitrogen) and resuspended in PK15 medium with the viability dye ToPro-3 (Invitrogen) at a concentration of 1–2·10^5 cells/ml. Live PK15 cells were single-cell sorted using a BD FACSAria II SORP UV (BD Biosciences) with 100 mm nozzle under sterile conditions. SSC-H versus SSC-W and FSC-H versus FSC-W doublet discrimination gates and a stringent '0/32/16 single-cell' sorting mask were used to ensure that one and only one cell was sorted per well. We sorted cells in 96-well plates with each well containing 100µl PK15 medium. After sorting, plates were centrifuged at 70g for 3 min. Colony formation was seen 7 days after sorting and we performed genotyping experiment 2 weeks after FACS.

To genotype single PK15 cells without clonal expansion, we directly amplified the PERV locus from sorted single cells according to a previously reported single cell genotyping protocol (*4*). Briefly, prior to sorting, we treated all plastics and non-biologic buffers with UV radiation for 30 min. We sorted single cells into 96-well PCR plates with each well carrying 0.5µl 10X KAPA express extract buffer (KAPA Biosystems), 0.1 µl of 1U/µl KAPA Express Extract Enzyme and 4.6 µl water. We incubated the lysis reaction at 75 ºC for 15 min and inactivated the reaction at 95 ºC for 5 min. All reactions were then added to 25µl PCR reactions containing 12.5µl 2X KAPA 2G fast (KAPA Biosystems), 100 nM PERV illumina primers (Methods Table2), and 7.5µl water. Reactions were incubated at 95 ºC for 3 min followed by 25 cycles of 95 ºC, 10 s; 65 ºC, 20 s and 72 ºC, 20 s. To add the Illumina sequence adaptors, 5µl of reaction products were then added to 20 µl of PCR mix containing 12.5 ml of 2 KAPA HIFI Hotstart Readymix (KAPA Biosystems), 100 nM primers carrying Illumina sequence adaptors and 7µl water. Reactions were incubated at 95 ºC for 5 min followed by 15-25 cycles of 98 ºC, 20 s; 65 ºC, 20 s and 72 ºC, 20 s. PCR products were checked on EX 2% gels (Invitrogen), followed by the recovery of 300-400bp products from the gel. These products were then mixed at roughly the same amount, purified (QIAquick Gel Extraction Kit), and sequenced with MiSeq Personal Sequencer (Illumina). We than analyzed deep sequencing data and determined the PERV editing efficiency using CRISPR-GA (*5*)**.**

Methods Table 2- Primers used in the PERV *pol* genotyping

Name	Sequence
illumina_primerPol1	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGACTGCCCCAAGGGTTCAA
illumina_primerPol2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCCGACTGCCCCAAGAGTTCAA
illumina_primerPo3	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTCTCTCCTGCAAATCTGGGCC

Targeting efficiency estimation: We built a custom pipeline to estimate the efficiency of PERV inactivation. Briefly, we amplified the *pol* gene and sequenced it via Illumina Next Generation Sequencing using PE250 or PE300. First, we combined the two overlapping reads using PEAR (*6*) and mapped to the reference region using BLAT. After mapping, we grouped the reads into sets containing specific combinations of haplotypes (see Extended Data Fig. 7), and indel types. Read sets with representation lower than 0.5% of the total number of mapped reads were discarded. Finally, the mapping output was parsed to call the different insertions and deletions as described in Güell et al (*5*) .

RNA-seq analysis: The susScr3 pig genome and Ensembl transcripts were obtained from the UCSC Genome Brower Database. RNA-Seq reads were mapped to the reference genome using the STAR software (*7*) and the RPKM of the transcripts were quantified using BEDTools (*8*). Differential expression analysis was performed in R using the DESeq2 package (*9*), and gene set enrichment analysis was carried out by the GSEA software (*10*), with gene set definitions obtained from the software's website.

Reverse transcriptase (RT) assay: To test the RT activity of the PK15 cells and modified PK15 clones (4 highly and 1 lowly modified clones), we plated $5 \cdot 10^5$ cells in T75 cm^2 flasks, and collected the supernatant 4 days after seeding. The media was filtered using a 0.45 µM Millex-HV Syringe Filter (EMD Millipore Corporation), and the filtered supernatant was concentrated at 4000g for 30min using Amicon Ultra-15 Centrifugal Filter Unit (EMD Millipore Corporation). The concentrated supernatant was ultra-centrifuged at 50,000 rpm for 60 min. The supernatant was carefully removed, and the virus pellet was collected and lysed with 20 µl of 10% NP40 at 37℃ for 60 min.

The RT reaction was conducted using the Omniscript RT Kit (Qiagen). The total volume of the reaction was 20 µl, which contained 1× RT buffer, 0.5 mM dNTPs, 0.5 µM Influenza reverse primer (5' CTGCATGACCAGGGTTTATG 3'), 100 units of RnaseOUT (Life Technology, Invitrogen), 100 units of SuperRnase Inhibitor (Life Technologies), 5 µl of sample lysis and 40 ng of IDT-synthesized Influenza RNA template which was rnase resistant in both 5' and 3' end. The RNA template sequence was 5' rA*rA*rC*rA*rU*rGrGrArArCrCrUrUrUrGrGrCrCrCrUrGrUrUrCrArUrUrUrUrArGrArArArUrCrArArGrUrCrArArGrArUrArCrGrCrArGrArArGrArGrUrArGrArCrArUrArArArCrCrCrUrGrGrUrCrArUrGrCrArGrArCrU*rC*rA*rG*rU*rG 3' (* phosphodiester bond). After the RT reaction was completed, the RT product was examined by PCR using Influenza forward (5' ACCTTTGGCCCTGTTCATTT 3') and Influenza reverse primers (sequence shown as above). The expected size of the amplicon was 72bp.

Infectivity assay

HEK293-GFP cell line establishment: The Lenti-GFP construct was derived from the plasmid pLVX-IRES-ZsGreen1 (Clontech. Catalog No. 632187; PT4064-5). To generate the lentivirus carrying Lenti-GFP, we transfected ~ $5 \cdot 10^6$ 293FT HEK cells with 3µg of pVX-ZsGreen plasmid and 12 µg of ViraPower Lentiviral Packaging Mix (Invitrogen) using Lipofectamine 2000 (Invitrogen). Lentiviral particles were collected 72 hours after transfection, and the viral titer was measured using Lenti-X GoStix (Takara Clonetech). We transduced ~ 10^5 lentivirus particles to ~ $1 \cdot 10^6$ HEK293 cells and conducted selection by puromycin to enrich the transduced cells 5 days after transduction. The 293-GFP-Lenti cell lines were maintained with 0.5 µg/mL puromycin thereafter.

Infectivity test of PK15 WT to HEK293-GFP: $1 \cdot 10^5$ cells of Lenti-GFP-293FT HEK cells and $1 \cdot 10^5$ PK15 WT cells were cultured together in a 6-well plate. In parallel, $2 \cdot 10^5$ PK15 WT cells were cultured alone in another well as a control. The puromycin selection experiment was done by adding 5 µg/ml of the antibiotic for 7 days. We determined the time point when no viable cells in the control well and approximately 100% GFP positive cells in the experimental well as the time point when we completed the puromycin selection to purify lenti-GFP-293FT human cells. Cells from the 293FT HEK/PK15 WT co-

culture were collected at different time periods. The genomic DNA was extracted using (DNeasy Blood & Tissue Kit, Qiagen) from cultured cells of the 293-GFP WT, PK15 WT and the co-cultured cells. The genomic DNA concentration was measured using a Qubit 2.0 Fluorometer (Invitrogen), and 3 ng from each sample was used as DNA template for PCR. In all, 1 µL of the genomic DNA were added to 25 µL of a PCR mix containing 12.5 µL 2X KAPA Hifi Hotstart Readymix (KAPA Biosystems) and 100 µM of primers as listed in Methods Table 3. Reactions were incubated at 95°C for 5 min followed by 35 cycles of 98°C, 20 s; 65°C, 20 s and 72°C, 20 s. PCR products were visualized on EX 2% gels (Invitrogen) and observed for bands of 300-400 base pairs.

Methods Table 3- A table exhibiting the primers used in the infectivity assay

Name	**Sequence**
PERV *pol* –Forward	GGG AGT GGG ACG GGT AAC CCA
PERV *pol* –Reverse	GCC CAG GCT TGG GGA AAC TG
PERV *env* –Forward	ACC TCT TCT TGT TGG CTT TG
PERV *env* –Reverse	CAA AGG TGT TGG TGG GAT GG
PERV *gag* –Forward	CGC ACA CTG GTC CTT GTC GAG
PERV *gag* –Reverse	TGA TCT AGT GAG AGA GGC AGA G
Pig GGTA1 –Forward	GGA GCC CTT AGG GAC CAT TA
Pig GGTA1 –Reverse	GCG CTA AGG AGT GCG TTC TA
Human ACTB–Forward	GCC TTC CTT CCT GGG CAT GG
Human ACTB–Reverse	GAG TAC TTG CGC TCA GGA GG

Quantification of PERV copy numbers infected in HEK293-GFP cells: We performed qPCR to quantify the PERV copy number in HEK293-GFP cells. Genomic DNA of PK15 WT cells of different amounts was used as the template for the qPCR reactions. Reactions were conducted in triplicate using KAPA SYBR FAST qPCR Master Mix Universal (KAPA Biosystems). PERV *pol*, *env*, *gag* primers, human ACTB and pig GGTA1 primers (Methods Table 3) were added to a final concentration of 1µM. Reactions were incubated at 95°C for 3 min (enzyme activation) followed by 50 cycles of 95°C, 5 s (denaturation); 60°C, 60 s (annealing/extension). We observed that the logarithm of the genomic DNA

amount linearizes with the quantification cycle (Cq). We used pol, gag, env primers to examine for presence of PERVs. Pig GGTA1 primers served to control for potential porcine genome contaminants in human cells after infection. All experiments were conducted in triplicate.

Infectivity Assay of the Modified PK15 clones to HEK293-GFP: $1 \cdot 10^5$ cells of HEK293-GFP cells and $1 \cdot 10^5$ cells of the high modified (15, 20, 29, 38) clones and low modified clones (40, 41) were co-cultured in a 6-well plate for 7 days. To isolate the HEK293-GFP cells in order to examine for PERV elements, we double sorted the GFP positive cells to purify the human cell populations.

To quantify the PERV infectivity of different clones to HEK293-GFP cells, we conducted both qPCR assays and PCR assays on series diluted HEK293-GFP cells after sorting. For the qPCR assays, we extracted the genomic DNA (DNeasy Blood & Tissue Kit, Qiagen) from double sorted HEK293-GFP cells. The genomic DNA concentration was measured using Qubit 2.0 fluorometer (Invitrogen). In all, 3 ng of the genomic DNA was added to 20 µL of KAPA SYBR FAST qPCR reaction (KAPA Biosystems) using PERV *pol*, *env*, *gag* and pig GGTA primers respectively (Extended Data Table 2). The qPCR procedure was performed as described above. For the series dilution assay, we further sorted purified HEK293-GFP cells (1 cell/ well, 10 cells/well, 100 cells/well, 1000 cells/well) into 96-well PCR plates for direct genomic DNA extraction and PCR reactions. Briefly, cells were sorted into 20 µL lysis reaction including 2 µL of 10X KAPA Express Extract Buffer, 0.4 µL of 1 U/µl KAPA Express Extract Enzyme and 17.6 µL of PCR-grade water (KAPA Biosystems). The reactions were then incubated at 55°C for 10 min (lysis), then at 95°C for 5 min (enzyme inactivation). Subsequently, the PCR master mix was prepared. In all, 2 µL of the genomic DNA lysis was added to 4 different 25 µL of KAPA Hifi Hotstart PCR reactions (KAPA Biosystems) using 1 µM PERV *pol*, *env*, *gag* primers, and pig GGTA primers, respectively (Extended Data Table 2). The reactions were incubated at 95°C for 3 min (initial denaturation) followed by 35 cycles of 95°C, 15 s (denaturation); 60°C, 15 s (annealing), 72°C, 15 sec/kb, then 75°C, 1 min/kb (final

extension). (KAPA Biosystems). The PCR products were visualized on 96 well E-Gel® Agarose Gels, SYBR® Safe DNA Gel (Invitrogen).

CRISPR-Cas9 off-target analysis: We obtained whole genome sequencing (WGS) data for PK15 (untreated cell line) and clone 20 (highly edited clone). To investigate potential off-target effects of the Cas9/2gRNAs, we first searched the reference sequence (*Sus Scrofa 10.2*) for sites that differed from the 20 bp sequences targeted by the two gRNAs by only 1 or 2 bp. We identified 11 such sites and extracted them, together with 200 bp of their neighboring regions (Fig. S1). We used BLAT to map the WGS reads to the extracted reference sequences and searched for potential indel patterns that had emerged in Clone 20 as a result of off-target effects. We obtained an average coverage of 7-8 X per loci. We excluded reads with <50 bp matches with the reference sequence. In case of reads that mapped to the reference sequence with multiple alignment blocks, which could indicate the presence of indels, we excluded reads whose alignment blocks contained <20 bp matches with the reference sequence. After inspecting the remaining mapped reads, we did not detect any off-target indel patterns present in clone 20. We should point out that another challenge to comprehensive searches for off-targets here is that the *Sus Scrofa* genome is still neither complete nor completely assembled, limiting our ability to do whole-genome analysis. We will be able to and fully intend to better search for off-targeting as these resources improve.

Mathematical model of DNA repair process interaction during cumulative PERV inactivation: In this study PERV elements were inactivated by mutations generated by DNA repair processes in response to dsDNA cuts created by Cas9. It is generally understood that dsDNA cuts may be repaired either by non-homologous end joining (NHEJ) or Homologous Repair (HR), and that while HR can create precise copies of a DNA template sequence at the cut site given the presence of a template with suitable homology arms, NHEJ can generate mutations (especially indels) and is often considered "error prone." However, there is also evidence that NHEJ can also repair dsDNA cuts highly accurately (*11*, *12*), and the relative rates of mutated vs. perfect repair by NHEJ have never been precisely measured. Especially when efficient targeted nucleases such as Cas9 are expressed for protracted time periods, perfect repair of a cut site by either

NHEJ or HR would regenerate a target site that could be cut again. A plausible hypothesis is that the process of perfect repair and re-cutting would occur repeatedly until a mutation arose that destroyed the nuclease's ability to recognize the target site. To explore the way these repair modalities might work together during the course of PERV elimination, we modeled their interactions as a Markov process. Specifically, we assumed:

i. There are N identical copies of the nuclease target in a cell.
ii. Only wild-type targets are recognized and cut, and only one target is cut and repaired at a time.
iii. DNA repair is either
    a. perfect restoration of the target site by NHEJ (with probability n)
    b. NHEJ that results in generation of a mutation that ablates target recognition (with probability m)
    c. repair by HR using any one of the other N−1 target sequences in the cell (with probability h)

We have n+m+h=1.

Our Markov model computes the probability distribution $\mathrm{P}^{(c)} = \left(\mathrm{p}_0^{(c)}, \mathrm{p}_1^{(c)}, \cdots, \mathrm{p}_N^{(c)}\right)$, where $\mathrm{p}_i^{(c)}$is the probability that there are i target-ablating mutations at cut c, where c = 0, 1, 2 … . We assume the initial condition $\mathrm{P}^{(0)} = (1,0,\cdots,0)$, i.e., that all targets begin as wild-type. The N+1-by-N+1 transition matrix M is given as

$$\left.\begin{aligned} \mathrm{M}(i,i) &= n + h \cdot \frac{N-i-1}{N-1} \\ \mathrm{M}(i,i+1) &= m + h \cdot \frac{i}{N-1} \end{aligned}\right\} \text{ for } 0 \le i < N$$

$$M(N,N) = 1 \qquad \text{for } i = N$$

$$M(i,j) = 0 \qquad \text{for all other } 0 \le i,j \le N$$

Finally, we have

$$P^{(c+1)} = P^{(c)}M \quad \text{for } c = 0,1,2,\cdots$$

The formulas for M assume proposition ii above and state in mathematical terms that the number of mutated sites in a cell remains unchanged whenever a cut at a wild-type site is repaired perfectly by NHEJ or by HR using another copy of the wild-type template (formula for $M(i,i)$), but increases by one if the cut is repaired by mutagenic NHEJ or by HR using a previously mutated site (formula for $M(i,i+1)$).

The model incorporates two notable simplifications to actual biology: (i) Target recognition is assumed to be binary – either the nuclease recognizes a target or it does not. This is tantamount to assuming that small mutations that still support target recognition do not substantially alter wild-type cutting rates and therefore can be effectively lumped together with wild-type sites. (ii) HR repairs using mutated vs. wild-type templates are assumed to be equally efficient. Modifications could be made to the model to address these simplifications, but this is not considered here. It is also worth noting that, formally, given assumption ii above, the Markov process should actually stop should the condition $p_N^{(c)} = 1$ be reached for some value of c, since at this point no wild-type sites remain to be cut, whereas what happens instead mathematically is that cuts continue but the model remains in a fixed state. Finally, the model effectively represents the mutation count distribution as a function of independent variable c (number of cuts) and not as a function

of time. No prediction is made regarding the time rates of DNA repair or PERV site elimination, although time can be assumed to increase monotonically with c.

To analyze PERV elimination through the Markov model, N was always set to 62. However, since the relative efficiencies of perfect vs. mutagenic NHEJ repair are unknown (as noted above), and because relative rates of mutagenic NHEJ vs. HR repair can vary widely depending on cell state and type, we computed mutation count distributions for a discrete grid covering the complete two-dimensional space of all possible parameter values for n, m, and h, (2500 parameter combinations in all). We implemented the model both as a MatLab (Mathworks, Waltham) script and as an R script using the library markovchain (*13*) (available as Supplemental Files modelMarkov.m, modelMarkov.R, respectively).

In addition to computing the mutation count distribution via the Markov model for particular parameter values, the MatLab script performed random simulations of the NHEJ and HR repair processes throughout a series of K cuts, allowing bivariate distributions of the numbers of total mutations vs. distinct NHEJ events to be estimated, illustrated in Fig. 27 B-C. The R script was used to estimate the most likely state of the system over the grid of n, m, and h combinations described above. K was varied depending on the computation. As illustrated in Fig. S27, the invariable result of the model was a unimodal distribution of mutation counts whose mean advanced towards fixation at N mutations with c, and in Figures S27 B.C, K was set to a value high enough to demonstrate fixation. For the calculation of the most likely state of the system over the n, m, and h grid, K was set to 50, 100, 200, or 500, and 100 simulations were conducted for each parameter combination.

Data deposition

Illumina Miseq data with PERVs elements genotyping data has been uploaded to the European Nucleotide Archive (ENA) hosted by the European Bioinformatics Institute (EBI) with the submission reference PRJEB11222.

References

1. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–7 (2004).

2. G. Wang *et al.*, Modeling the mitochondrial cardiomyopathy of Barth syndrome with induced pluripotent stem cell and heart-on-chip technologies. *Nat. Med.* **20**, 616–23 (2014).

3. N. E. Sanjana, O. Shalem, F. Zhang, Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods*. **11**, 783–784 (2014).

4. G. D. Evrony *et al.*, Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell*. **151**, 483–96 (2012).

5. M. Güell, L. Yang, G. M. Church, Genome editing assessment using CRISPR Genome Analyzer (CRISPR-GA). *Bioinformatics*. **30**, 2968–2970 (2014).

6. J. Zhang, K. Kobert, T. Flouri, A. Stamatakis, PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*. **30**, 614–20 (2014).

7. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

8. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

9. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*, 1–21 (2014).

10. A. Subramanian *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–50 (2005).

11. S. M. Byrne, L. Ortiz, P. Mali, J. Aach, G. M. Church, Multi-kilobase homozygous targeted gene replacement in human induced pluripotent stem cells. *Nucleic Acids Res.* **43**, e21 (2014).

12. M. Bétermier, P. Bertrand, B. S. Lopez, Is non-homologous end-joining really an inherently error-prone process? *PLoS Genet.* **10**, e1004086 (2014).

13. CRAN - Package markovchain, (available at https://cran.r-project.org/web/packages/markovchain/).