Title: The HERV-K HML-2 integration within *RASGRF2* is associated with intravenous drug abuse and modulates transcription in a cell-line model

+Correspondence should be addressed to:

aris.katzourakis@zoo.ox.ac.uk

gmagi@med.uoa.gr

This file includes:

Materials and Methods

Supplementary Figures

Other Supplementary Materials for this manuscript includes the following:

Dataset S1

Materials and Methods

Wet Lab

Design of target enrichment and high throughput sequencing

We developed a custom library-preparation method for the targeted enrichment of HK2 LTR and their flanking sequences. To do this, we designed custom 120bp-long biotinylated probes from IDT to perform a capture of the first and the last ~360bp of the HK2-LTR sequence (5 probes overlapping by 60bp at each LTR side). The libraries were prepared with a target insert size at library Quality Control (QC) of 1200bp. From trial runs, we knew that a library QC showing an insert of approximately 1200bp would produce sequencing data with a considerably shorter insert. The prepared libraries were validated and multiplexed before 2X300bp paired-end sequencing on an Illumina MiSeq instrument.

Cells and reagents

General cell culture reagents were purchased from Sigma-Aldrich (Poole, U.K.) The human embryonic kidney HEK 293 cell line was purchased for this specific project from the European Collection of Authenticated Cell Cultures (ECCAC; Porton Down, U.K.) and the teratocarcinoma cell line NCCIT was obtained from the American Type Culture Collection (ATCC, LGC Standards, Teddington, U.K.). Both of the cell lines came with quality certificates for their identity and being free from mycoplasma contamination.

The cells were incubated in a humidified atmosphere at 37°C with 5% CO_2 and cultured as described below.

HEK 293 cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) with high glucose (4500mg/dL), L-glutamine and sodium bicarbonate. It was supplemented with 10% heat-inactivated foetal bovine serum, 1% penicillin-streptomycin and 2mM L-glutamine. The cells were allowed to grow to 100% confluence in a T75 flask and were typically subcultured twice per week using trypsin-EDTA to dislodge them from the substratum. The cell suspension was diluted 1:20 into a new flask. Two days after seeding a flask, the media was replaced and the spent media filtered through a 0.2μM syringe filter. Aliquots of the filtered media were stored at -80°C until needed for use as conditioned media in genome engineering.

The NCCIT cells were cultured in RPMI growth medium supplemented with 10% v/v foetal bovine serum, 1% v/v penicillin-streptomycin and 2mM L-glutamine. Two days after seeding the flask, the media was replaced with fresh growth media due to the elevated respiration rate. They were allowed to grow to 100% confluence in a T75 flask and were typically subcultured every four days using trypsin-EDTA to dislodge them from the substratum. On each passage, approximately $1x10^6$ cells were reserved for later gDNA extraction.

SCR7 was obtained from Stratech Scientific Limited (Newmarket, U.K.). Puromycin was purchased from Cayman Chemicals (Cambridge Bioscience, Cambridge, U.K.)

Genome Engineering

-CRISPR/Cas9 plasmids

The Cas9 and sgRNA plasmids were designed by Andrew Bassett and Joey Riepsaame of Genome Engineering Oxford (GEO) to introduce the HERV sequence into the RASGRF2 gene. We designed the HDR plasmid and this was synthesised using GeneArt (Thermo Fisher Scientific). Subsequently, for the HAP1 transfection experiments we sub-cloned the HDR construct into a new backbone to remove the CMV promoter and increase targeting efficiencies. This construct was converted into a minicircle (Supplementary Figure 3).

-Transfection of cell-lines with CRISPR/Cas9 plasmids

HEK 293 cells were seeded into a 6-well plate at a density of $1x10^6$ cells/well. The following day, the cells were co-transfected with 1µg each of the sgRNA and HDR plasmids using calcium phosphate. The transfection mixture was left on the cells overnight and then replaced with fresh growth media. The cells were allowed to recover for 24h, then the selection agent (puromycin) was added. Selection was conducted for 7d, with the media (containing selection agent) replaced every 1-2d. The cells were then examined by PCR for the presence of the insertion.

Subsequently, cells that were positive for the integration were subjected to limiting dilution in order to create clonal cultures. Briefly, the cells were detached using trypsin, counted and a dilution made to produce a concentration of 100 cells/mL in 80% normal growth media and 20% conditioned media. These cells were then used to seed a 96 well plate (100µL/well). The next day, the plate was examined and wells containing a

single cell were marked. The plate was then returned to the incubator and monitored for 7-10 days. When the positive wells reached a high density, they were subcultured and used to seed cluster plates with larger wells. This was repeated until the cells were growing in 6wp. Finally, once the density of the 6-well plate was high enough, some cells were removed for gDNA extraction (Qiagen DNeasy Blood, Cells and Tissue kit) and PCR analysis.

A refinement of the above method involved using the NHEJ inhibitor SCR7. Briefly, the cells were seeded as above for transfection. The following day, the cells were treated with SCR7 2h before transfection. Importantly, the SCR7-containing media was not removed prior to the transfection. The transfection procedure was then conducted as described above.

- *CRISPR/Cas9 editing of eHAP1 cells*

Since HEK293s select the WT alleles over those with the RASGRF2-int, we tried to edit a haploid cell line (eHAP1), using the same HDR synthesised plasmid as for the HEK293s. For the integration screening, we used two different PCRs. Each consisted of a primer binding within the newly introduced DNA, the other binding outside (recognizing the WT sequence). We succeeded in screening individually picked and expanded clones but unfortunately, none of these survived upon expansion to have integrated the HDR construct (Supplementary Figure 3). In our 2nd attempt, we first sub-cloned the HDR construct into a new backbone to remove the promoter and increase targeting efficiencies. This construct was converted into a minicircle, devoid of any bacterial backbone (improves HDR efficiencies) (more info at: www.systembio.com/products/gene-expression-

systems/minicircle-technology/) eHap1 cells were targeted once more, using 200.000 cells - 5 ug Cas9 protein - 1 ug sgRNA - 1 ug minicircle DNA. We Nucleofected the cells following the Horizon Discovery protocol (1575V - 10 ms - 3 pulses). After nucleofection, cells were allowed to recover for 48h and then seeded at 5000 cells per 10 cm petri dish to grow into individual colonies. The remainder of cells (bulk) was used for triple genotyping PCR analysis. Unfortunately, the screening of the bulk DNA showed no sign of integrated DNA

-Controls of normal RASGRF2 transcription

We used as normal controls for the RASGRF2 the unedited HEK293 as well as edited HEK293s that "lost" the RASGRF2-int 48-days post-transfection (see also Figure 1). The first one is a "wild type" control, while the second one is a control that shows that transcription of RASGRF2 is restored when the integration is lost.

PCR screening for RASGRF2-int patient samples and qPCR on cell-lines

To estimate the prevalence of the RASGRF2-polymorphic integration in PWIDs and non-PWIDs, we designed two pairs of primers (one common reverse), which specifically amplify either the allele with the integration alone (primers LTR_splc_F - dn_intg_R) or both the wild type and the integrated alleles, giving distinct PCR products (primers up_intg_F - dn_intg_R) (Dataset S1). This method allows the screening and the genotyping of multiple samples simultaneously (Figure 1, a-d).

We further investigated the impact of the HK2-LTR integration on the expression of the host RASGRF2 gene. We used CRISPR/Cas9 technology to edit the HEK293 cell line and to manually insert the LTR sequence in the pre-integration intronic site between exons 17 and 18 (see above). Before proceeding with the CRISPR/Cas9 engineering experiments we genotyped our 293s cell line (ECACC 85120602) and we found it negative for the RASGRF2 LTR integration. Using publicly available data from *www.hek293genome.org/v1/data.php* we retrieved the rs26907 SNP genotypes for most of the available 293s derivatives and we found them heterozygous. In addition, we did not find rs26907 in published Neanderthal genomes whereas RASGRF2-int has been identified, while the frequency of rs26907 in East Asian populations is similar to other populations, where RASGRF2-int has not been identified (see main text). These observations do not indicate linkage between rs26907 and RASGRF2-int..

The LTR and the flanking host sequences were previously confirmed with Sanger step-sequencing in patient samples positive for the integration. We confirmed the integration and we genotyped the cell line in multiple post-clonal-selection time-points using allele-specific PCR designs. The edited allele was present in the heterozygous state in the selected clones, as the wild-type allele-specific PCR was also positive after the transfection. The forward primer (intg_conf_F) of the primer-set used for the confirmation of the integration, was annealing upstream to the left homologous arm used in the HDR plasmid, allowing us to distinguish the edited alleles from the HDR plasmid (Figure 1).

The expression levels of RASGRF2 exons were evaluated using multiple SYBR-Green qPCRs (primers RAS2-3, 16-17, 17-18, 18-19). We found decreased expression levels of exons around the integration, which were reversed to normal after the selection of the wildtype alleles, 48 days post-transfection (T75 flask).

Assessment of HK2 LTR promoter in plasmids

The HERVK-113 LTR *in vitro* synthesized dsDNA fragment (Invitrogen) was cloned into the HincII restriction site of pUC19 resulting HK-113-LTR/pUC19 recombinant plasmid. The HK-113-LTR XhoI excised fragment from HK-113-LTR/pUC19 plasmid was cloned into the XhoI restriction site of pGEM-luc vector (Promega) both in forward and reverse orientation resulting plasmids HK-113-LTR-Luc/F, HK-113-LTR-Luc/R respectively.

The human cell lines Huh7 (hepatocarcinoma) and HeLa (cervical carcinoma) were each maintained in Dulbecco's modified Eagle medium supplemented with 2 mM glutamine, 10% (v/v) heat-inactivated fetal calf serum and 100 U/ml penicillin/streptomycin.

For DNA transfections, Huh7 and HeLa cells were transiently transfected using JetPEI reagent (PolyPlus) according to the manufacturer's protocol. Following transfection, cells were harvested, seeded into 48-well plates, collected at the appropriate time points and subjected to luciferase assays with a commercially available kit (Promega). Luciferase activity was normalized to total cell protein in order to yield relative luciferase activity (RLA).

Huh7 and HeLa cells were transfected with pGEM-luc vector (control) and HK-113-LTR Luc/F, HK-113-LTR-Luc/R plasmids. 48h post-transfection, cells were lysed and luciferase activity was measured; luciferase readouts were normalised to total cell protein.

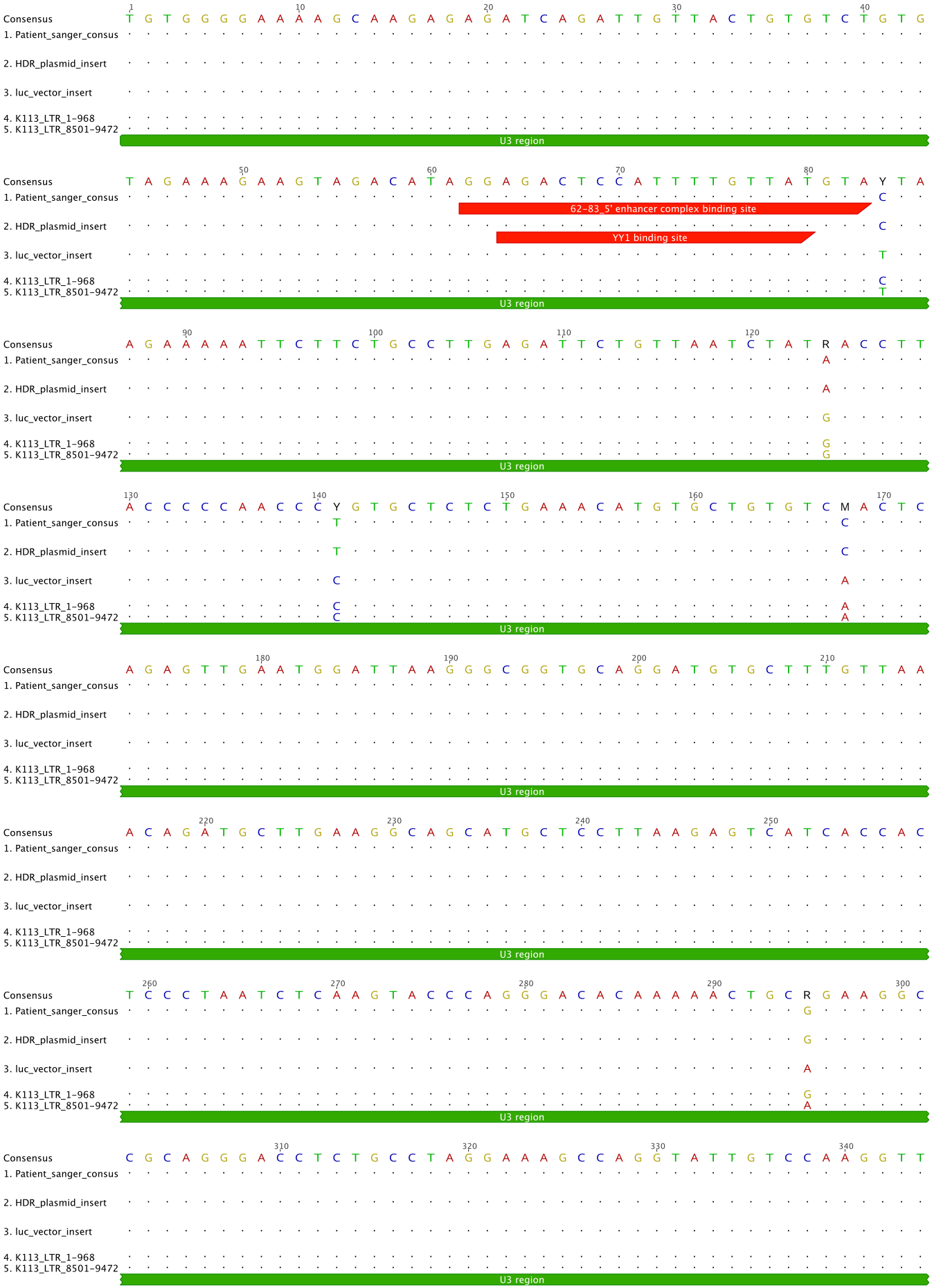
Expression patterns of read-through transcripts

We found significantly increased expression levels over the LTR-ends compared to the LTR-starts (Wilcoxon, $p<0.05$) (Figure 2). We also found that elements with increased expression levels over the LTR-start would present decreased expression (or would be inactive) over the LTR-end and vice-versa. (Figure 2)

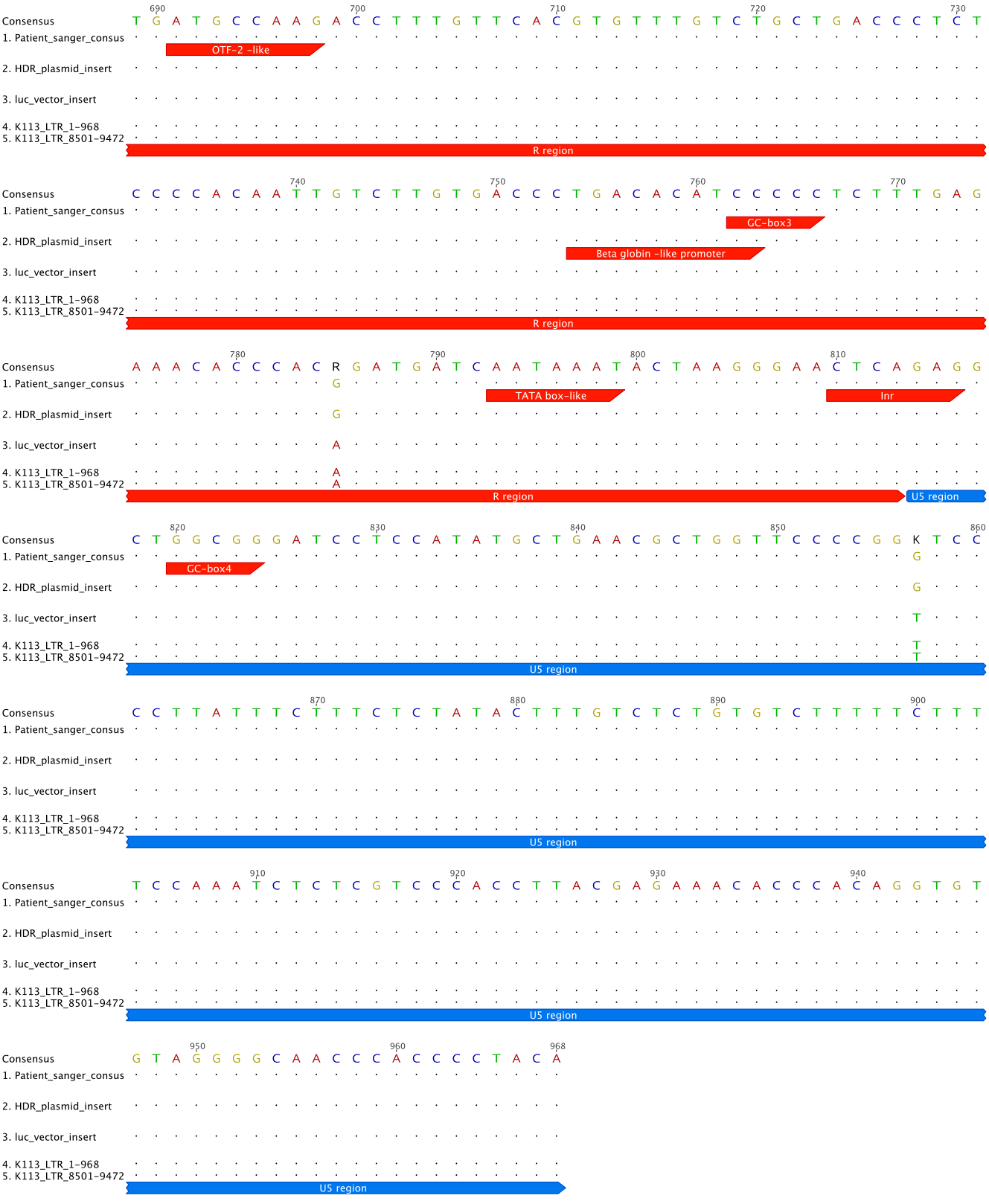
We tested the expression levels in combination with the type but also the estimated age of each element as described in Subramanian et al (1). We found that full-length elements are significantly more active compared to the solo-LTR integrations, over their LTR-end edges (Wilcoxon, $p<0.05$) (Figure 2). We asked for at least 5 reads per fragment to evaluate a HK2-edge as active and we estimated their expression (active vs. not-active) and the magnitude (<500 vs. >500 reads) per edge, in combination with their age (younger than 5my, 5-20my and older than 20my). We found that the youngest elements (younger than 5my) are more likely to be active versus the very old ones (older than 20my) over the LTR-ends. The same trend was observed for the magnitude of the expression, with the youngest elements to be more expressed versus the older ones, both over the LTR-start and the LTR-end regions (z-test, one-tailed, $p<0.05$) (Figure 2).

Supplementary Figures

Supplementary Figure 1: Nucleotide alignment and transcription regulatory motifs annotation of: (1) Sanger step sequencing consensus of the RASGRF2 – integrated solo LTR. (2) Synthetic fragment inserted into the HDR plasmid, which was used for the CRISPR/Cas9-mediated editing of the HEK293 cells (identical to 1). (3) Synthetic fragment used for the Luciferase expression vector construction. (4, 5) HERV-K 113 LTRs, regions: 1-968bp and 8501-9472bp, respectively. The K113 LTR sequences only differs by <1% compared to the RASGRF2 solo LTR integration, while none of the observed sequence differences is located within promoter/regulatory sites, e.g. the four GC-boxes which are targets of the transcription factors Sp1 and Sp3 (2), the transcription initiation site (Inr), the TATA boxes -like and the Beta globin distal promoter -like sequence or the binding site of the transcription factor OTF-2 (3) or the 5′-enhancer complex binding site (nt 62-83) and the YY1 transcription factor (nt 64-80) (4). The U3, R and U5 LTR regions are indicated in green, red and blue arrows respectively.


Consensus
1. Patient_sanger_consus
2. HDR_plasmid_insert
3. luc_vector_insert
4. K113_LTR_1-968
5. K113_LTR_8501-9472
62-83_5' enhancer complex binding site
YY1 binding site
U3 region


Consensus
1. Patient_sanger_consus
2. HDR_plasmid_insert
3. luc_vector_insert
4. K113_LTR_1-968
5. K113_LTR_8501-9472
U3 region
TATA box-like
GC-box1
GC-box2
R region



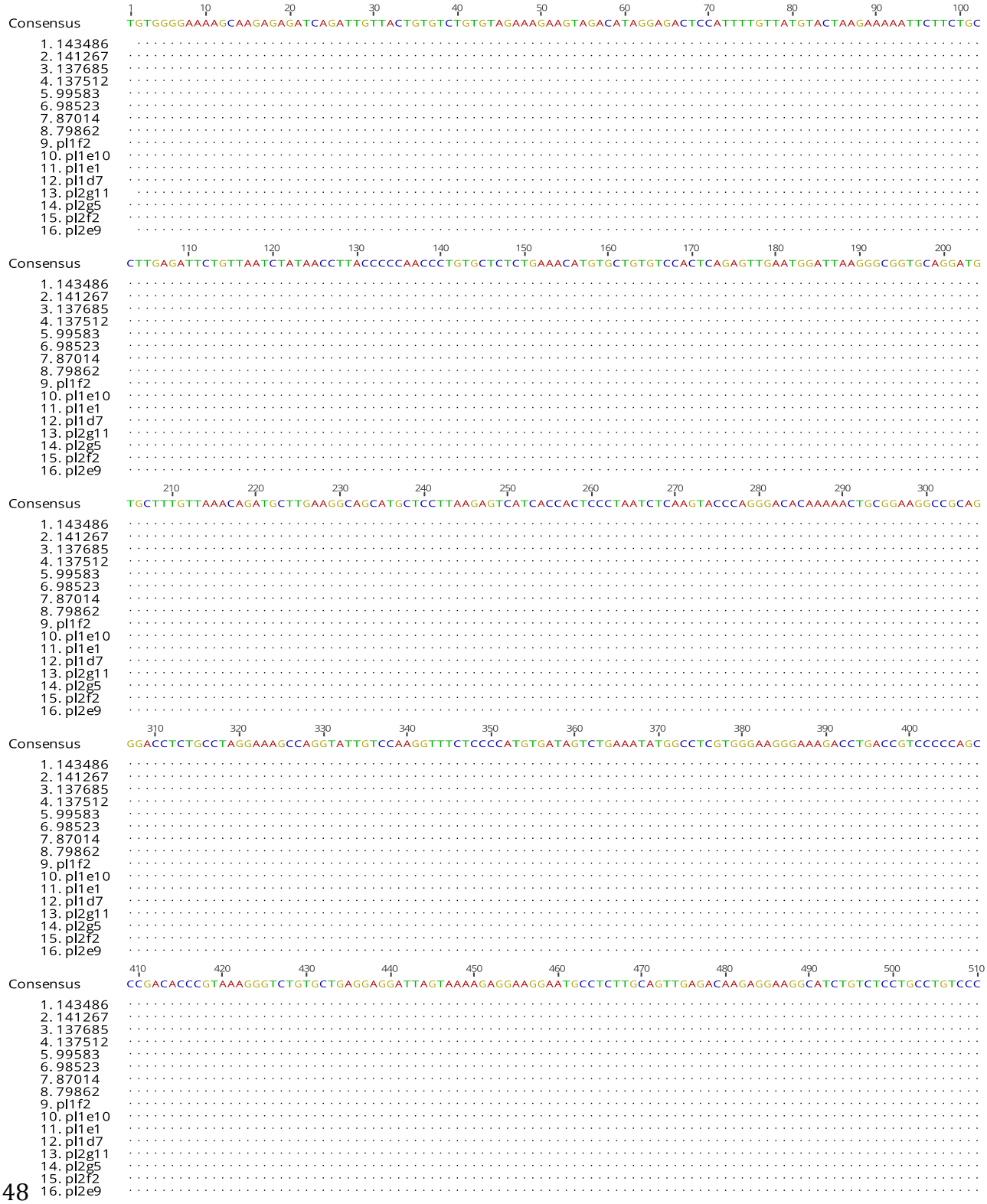
238

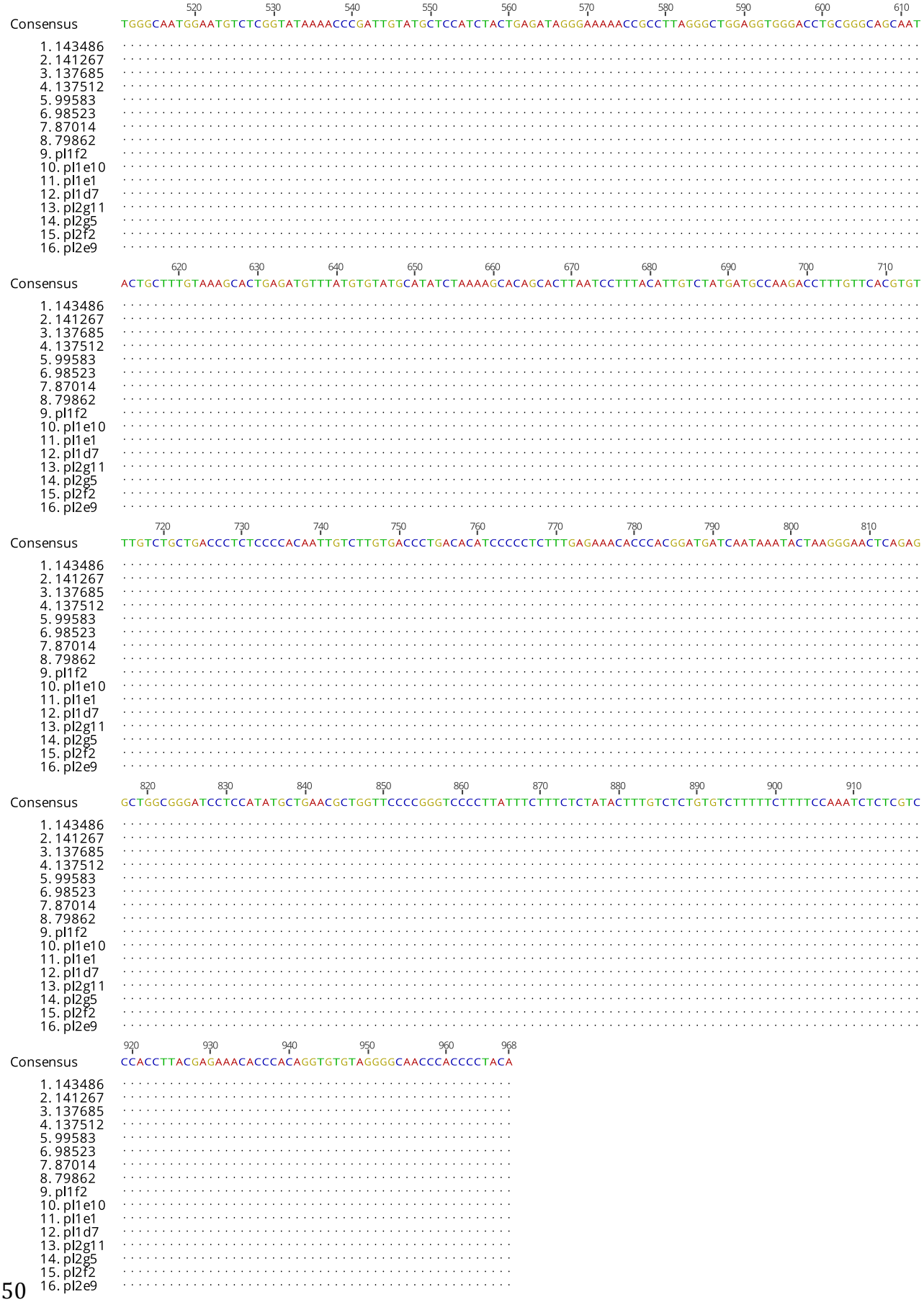
239

240

241

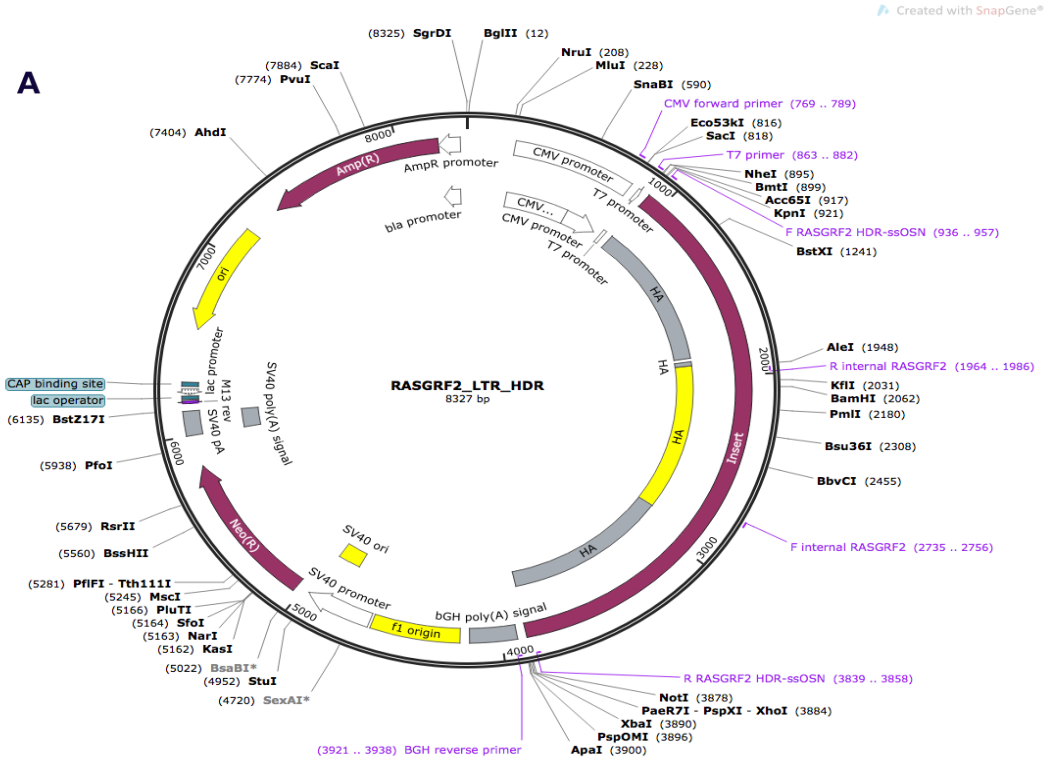
Supplementary Figure 2: Nucleotide alignment of 16 sanger-sequenced RASGRF2-int LTR positive patients, 4 from each group: 1-4: GR PWID (Greek Persons Who Inject Drugs),5-8: GR non-PWID (Greek non-PWID), 9-12: UK PWID (British Persons Who Inject Drugs),13-16: UK non-PWID. All samples were found identical.

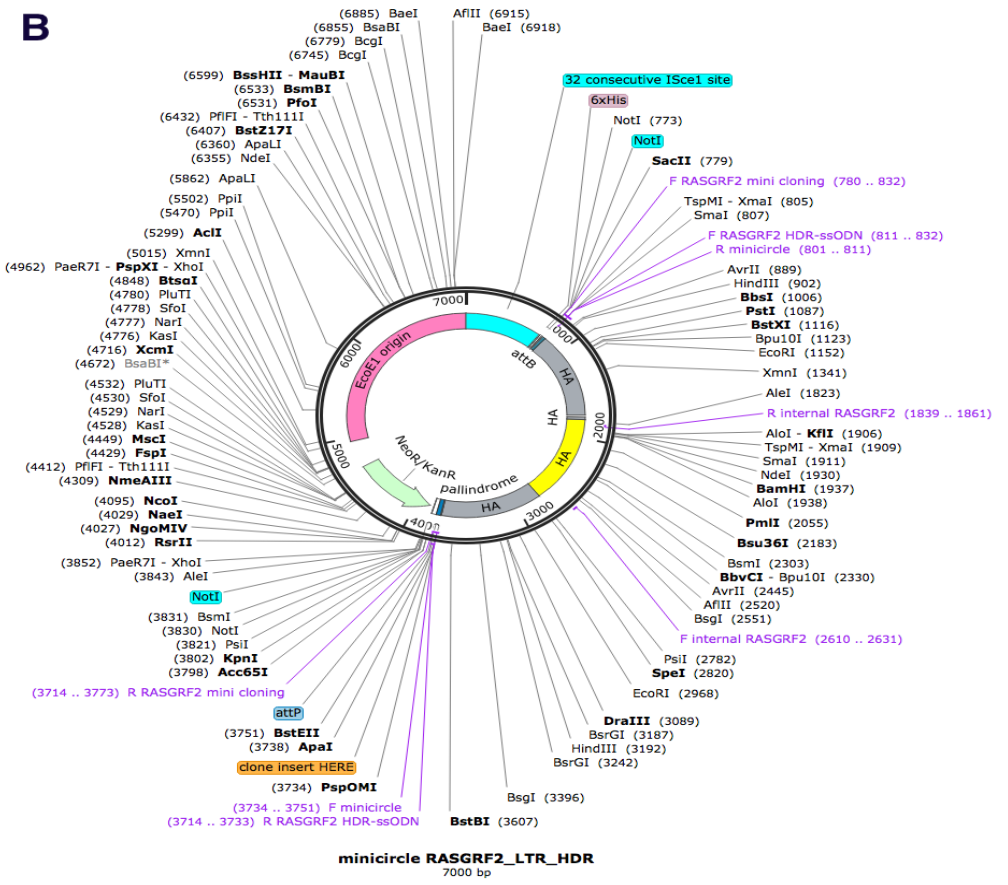



520 530 540 550 560 570 580 590 600 610
Consensus TGGGCAATGGAATGTCTCGGTATAAAACCCGATTGTATGCTCCATCTACTGAGATAGGGAAAAACCGCCTTAGGGCTGGAGGTGGGACCTGCGGGCAGCAAT
620 630 640 650 660 670 680 690 700 710
Consensus ACTGCTTTGTAAAGCACTGAGATGTTTATGTGTATGCATATCTAAAAGCACAGCACTTAATCCTTTACATTGTCTATGATGCCAAGACCTTTGTTCACGTGT
720 730 740 750 760 770 780 790 800 810
Consensus TTGTCTGCTGACCCTCTCCCCACAATTGTCTTGTGACCCTGACACATCCCCCTCTTTGAGAAACACCCACGGATGATCAATAAATACTAAGGGAACTCAGAG
820 830 840 850 860 870 880 890 900 910
Consensus GCTGGCGGGATCCTCCATATGCTGAACGCTGGTTCCCCGGGTCCCCCTTATTTCTTTCTCTATACTTTGTCTCTGTGTCTTTTTCTTTTCCAAATCTCTCGTC
920 930 940 950 960 968
Consensus CCACCTTACGAGAAACACCCACAGGTGTGTAGGGGCAACCCACCCCTACA
1. 143486
2. 141267
3. 137685
4. 137512
5. 99583
6. 98523
7. 87014
8. 79862
9. pl1f2
10. pl1e10
11. pl1e1
12. pl1d7
13. pl2g11
14. pl2g5
15. pl2f2
16. pl2e9

250
251

Supplementary Figure 3: Plasmid map of (A) the HDR vector used for the transfection of HEK293s and the first attempt of eHAP1s and (B) the minicircle construct used for the second transfection attempt of eHAP1s.

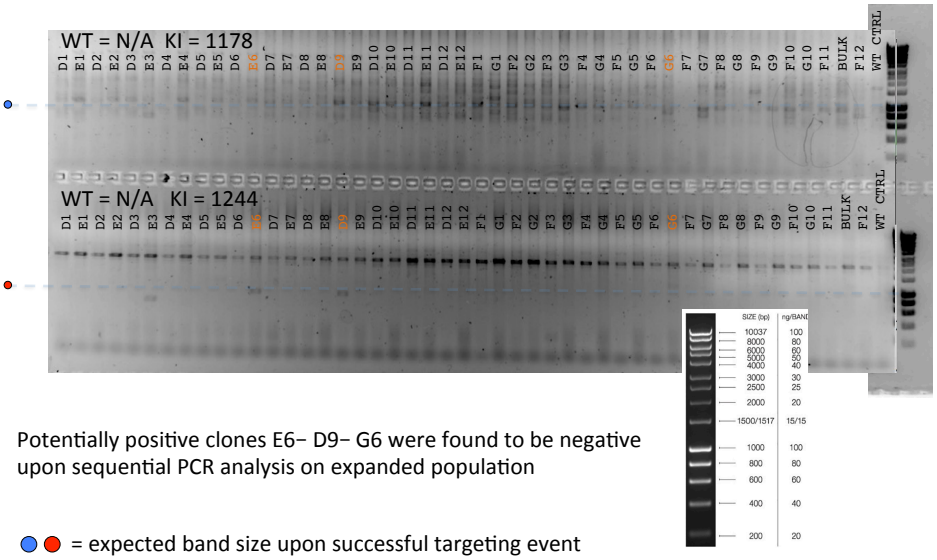




Supplementary Figure 4: CRISP\Cas9 editing of eHAP1 cells using (A) the same HDR vector as for the HEK293s and (B) the minicircle sub-construct

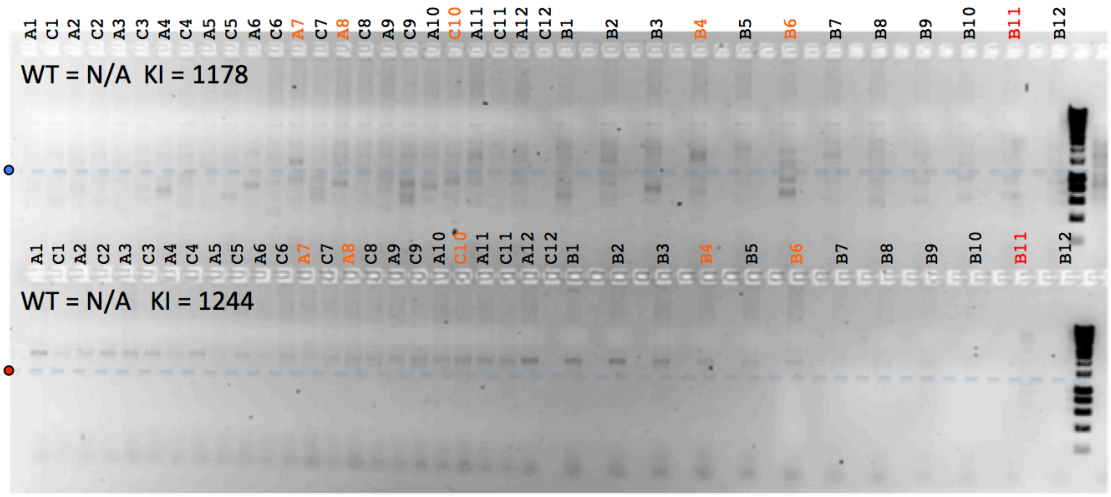
(A) RASGRF2 HERVK Integration (KI) PCR (KAPA2G 4kb) – 1st round

Conditions: 1 ul gDNA per 25 ul PCR – 1.25 ul of each 10 uM primer
Performed two different PCRs per sample



Potentially positive clones E6– D9– G6 were found to be negative upon sequential PCR analysis on expanded population

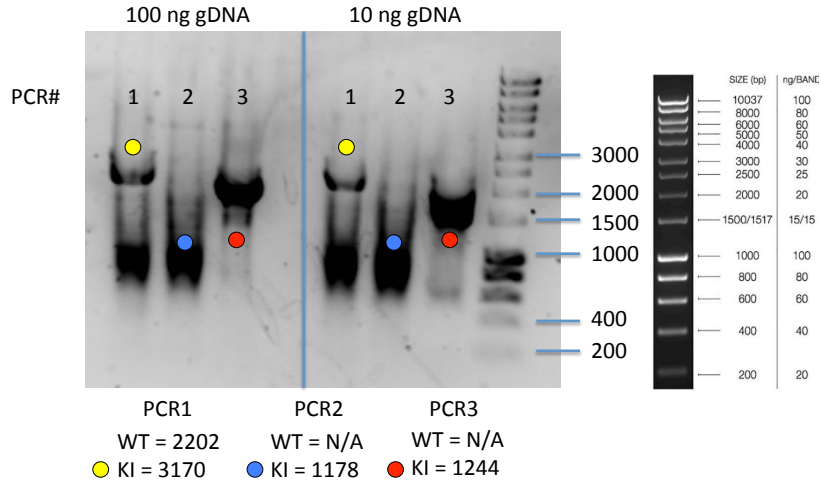
= expected band size upon successful targeting event



Potentially positive clones A7 – A8 – B4 – B6 – B11 – C10 were found negative upon sequential PCR in expanded population.

(B) RASGRF2 HERVK Integration (KI) PCR (KAPA2G 4kb) – 2nd round (minicircle)

Conditions: 1 ul gDNA (bulk of targeted cells) per 25 ul PCR – 1.25 ul of each 10 uM primer
Tm = 60 deg



= expected band size upon successful targeting event

References

1. Subramanian RP, Wildschutte JH, Russo C, & Coffin JM (2011) Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* 8.
2. Fuchs NV, *et al.* (2011) Expression of the human endogenous retrovirus (HERV) group HML-2/HERV-K does not depend on canonical promoter elements but is regulated by transcription factors Sp1 and Sp3. *J Virol* 85(7):3436-3448.
3. Kovalskaya E, Buzdin A, Gogvadze E, Vinogradova T, & Sverdlov E (2006) Functional human endogenous retroviral LTR transcription start sites are located between the R and U5 regions. *Virology* 346(2):373-378.
4. Knossl M, Lower R, & Lower J (1999) Expression of the human endogenous retrovirus HTDV/HERV-K is enhanced by cellular transcription factor YY1. *J Virol* 73(2):1254-1261.