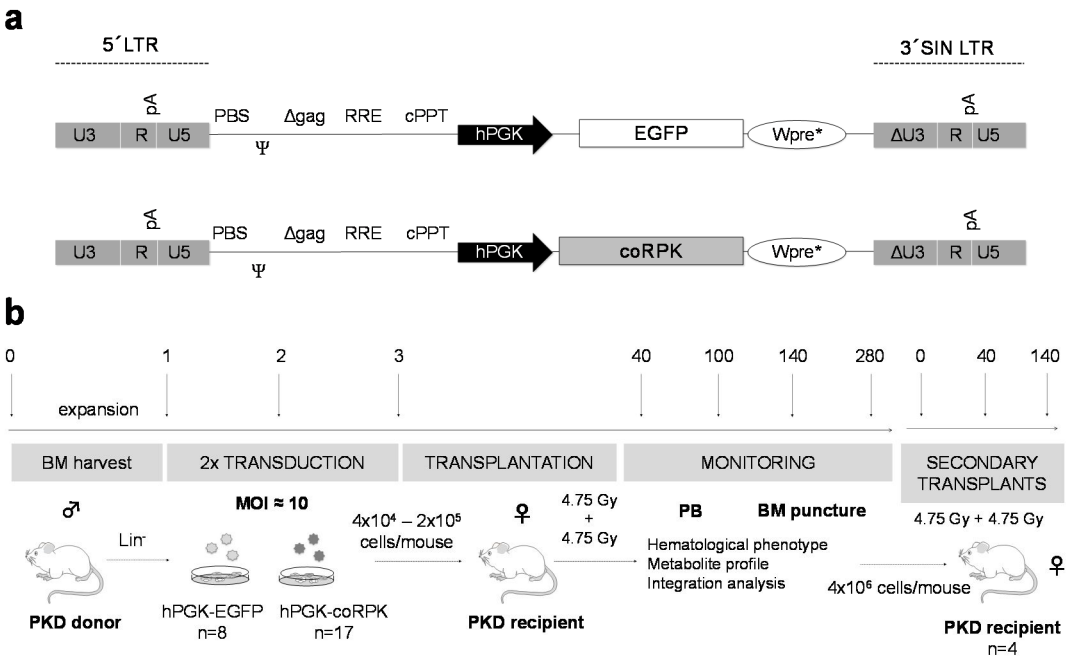


Figure S1. Experiment design. (a) Schematic representation of the self-inactivating lentiviral vectors used throughout gene therapy experiments harboring the human PGK promoter regulating the expression of the EGFP transgene in the control vector (upper diagram) or the expression of a codon-optimized sequence of the *PKLR* gene cDNA (coRPK) in the therapeutic vector (lower diagram). The coRPK sequence showed 80.4% homology with the human *PKLR* cDNA and 76.5% homology with the mouse *Pklr* cDNA, with no changes in the amino acid sequence. **(b)** Schematic gene therapy protocol performed to address the functionality of the developed PGK-coRPK lentiviral vector. Correction of the PKD phenotype was studied for 4 to 9 months after transplant in PB and BM through hematological analysis and metabolic profiling. Integration analysis was performed in different tissues and time-points from all mice to address LV vector safety. At 280 days post-transplantation, total BM from primary transplanted mice carrying the coRPK transgene was transplanted again into lethally-irradiated female PKD mice (secondary recipients) to test the stability and safety of the engraftment.

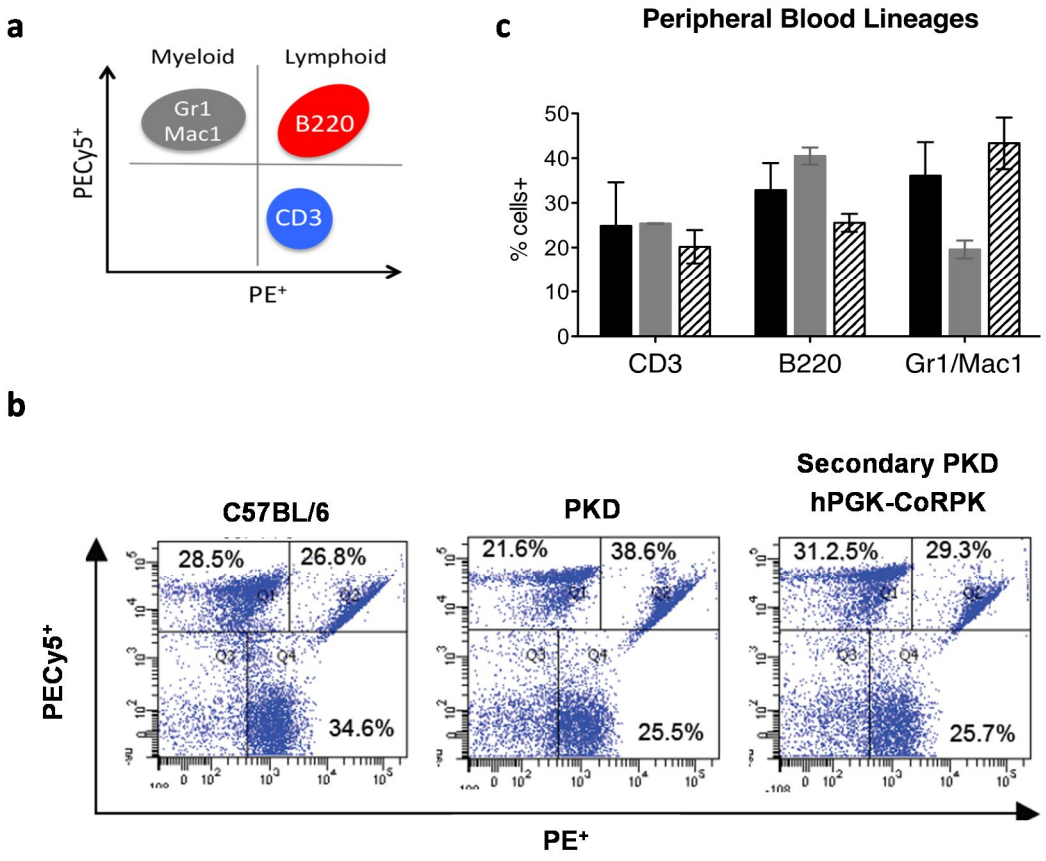


Figure S2. Multi-lineage hematopoietic reconstitution in secondary transplanted mice. (a) Diagram of the flow cytometry strategy used to identify the different hematopoietic lineages by labeling with CD3-PE, B220-PE, B220-PECy5, Gr1-Biotin and Mac1-Biotin antibodies plus SAV-PE-Cy5. **(b)** Representative dot-plots and **(c)** percentages of each lineage in PB at 140 days after transplant. Bars represent the average percentage ± SEM of healthy (n=2, black bar) and PKD mouse (n=2, grey bar) controls and secondary transplanted mice expressing the coRPK therapeutic transgene (n=4, scratched bar).

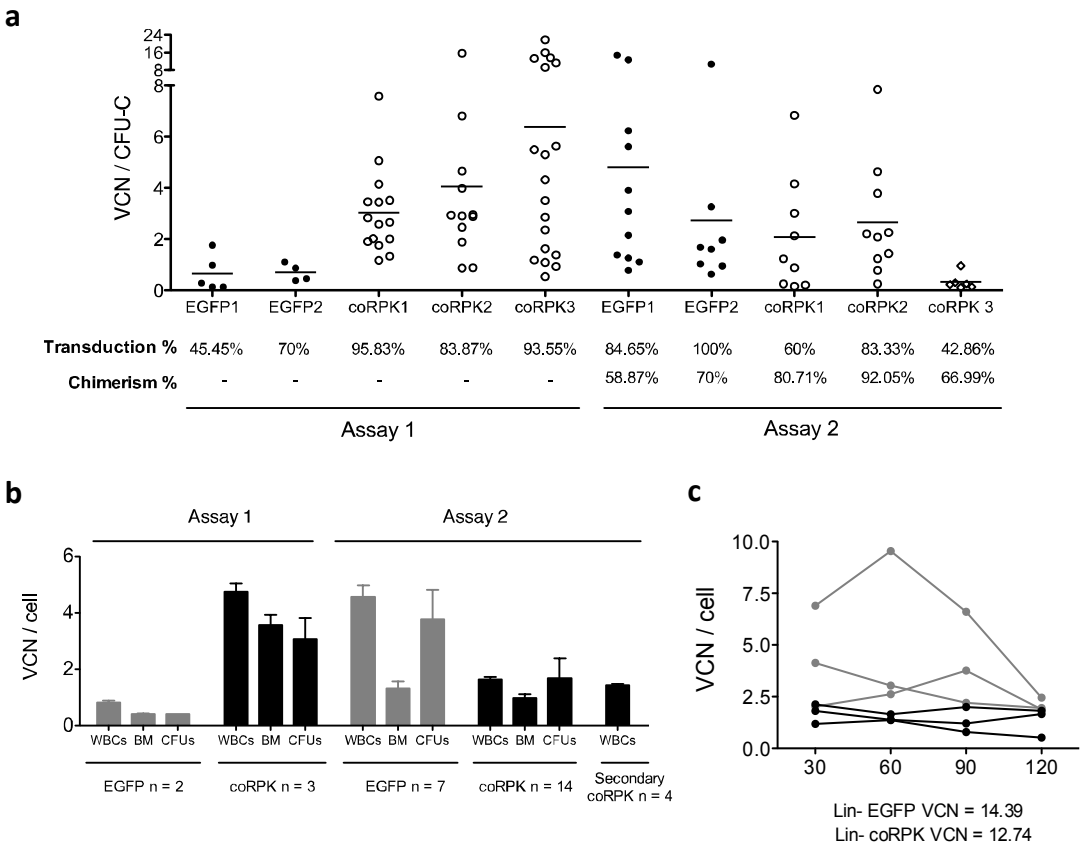


Figure S3. Quantification of proviral integrations. (a) Vector copy number per cell in BM CFUs from individual transplanted mice at 120 to 170 days after transplant. Transduction and chimerism percentages are also shown. **(b)** Provirus copy number in cells from different hematopoietic compartments. Columns represent the average ± SEM of the different groups of transplanted mice. **(c)** Kinetics of proviral integrations in BM cells from individual transplanted EGFP-expressing mice (grey lines) and mice carrying the coRPK transgene (black lines).

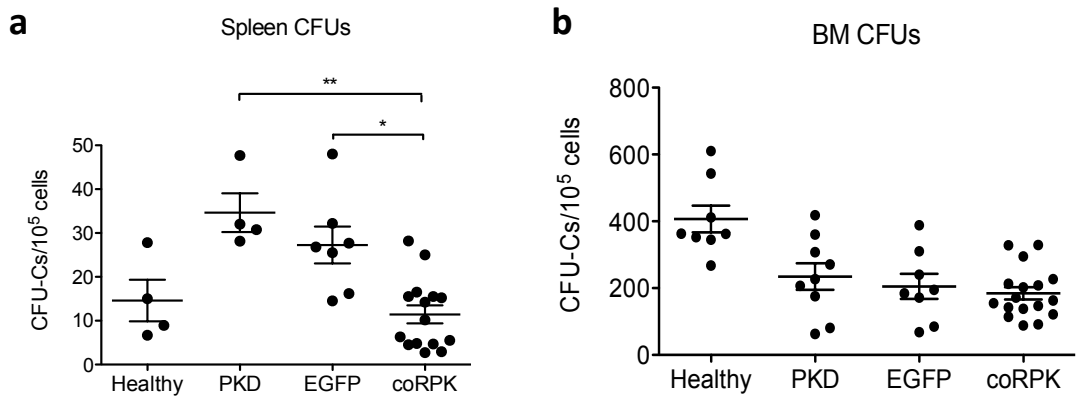


Figure S4. Hematopoietic progenitor assays in control mice and transplanted mice with transduced cells. (a) Total CFUs from spleen and **(b)** bone marrow at 140 days after transplant. Dots represent number of colonies per mouse analyzed and lines represent average ± SEM in each group. Data were statistically analyzed by non-parametric Kruskal-Wallis test.

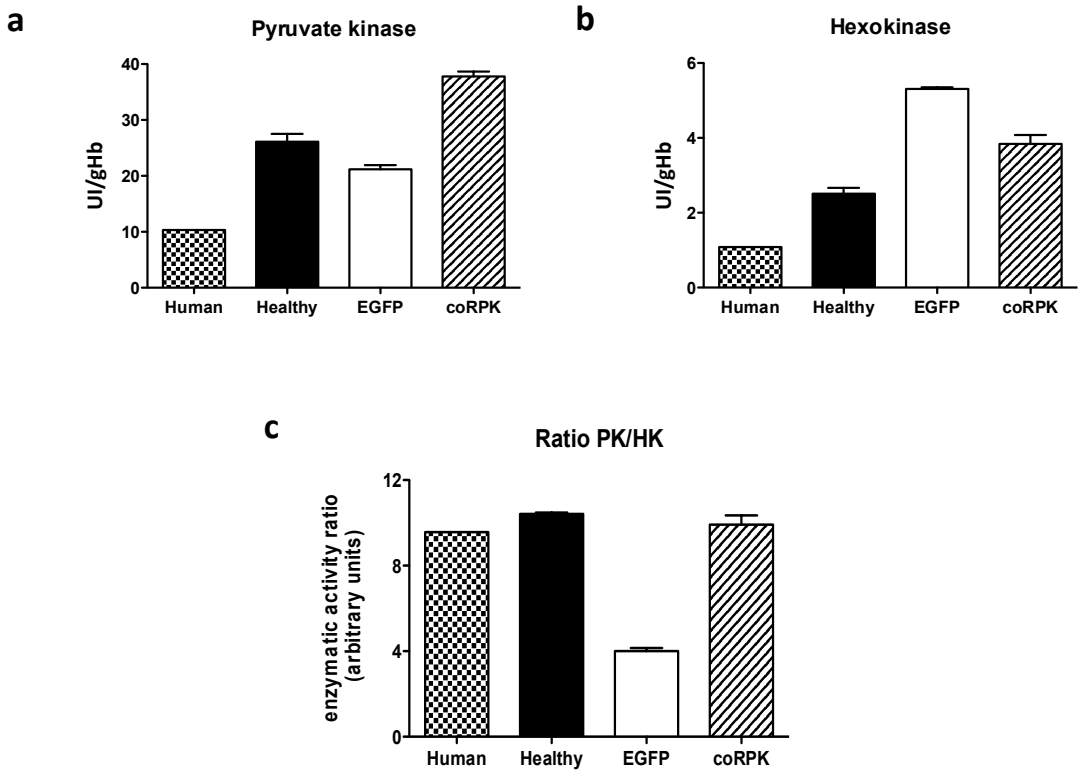
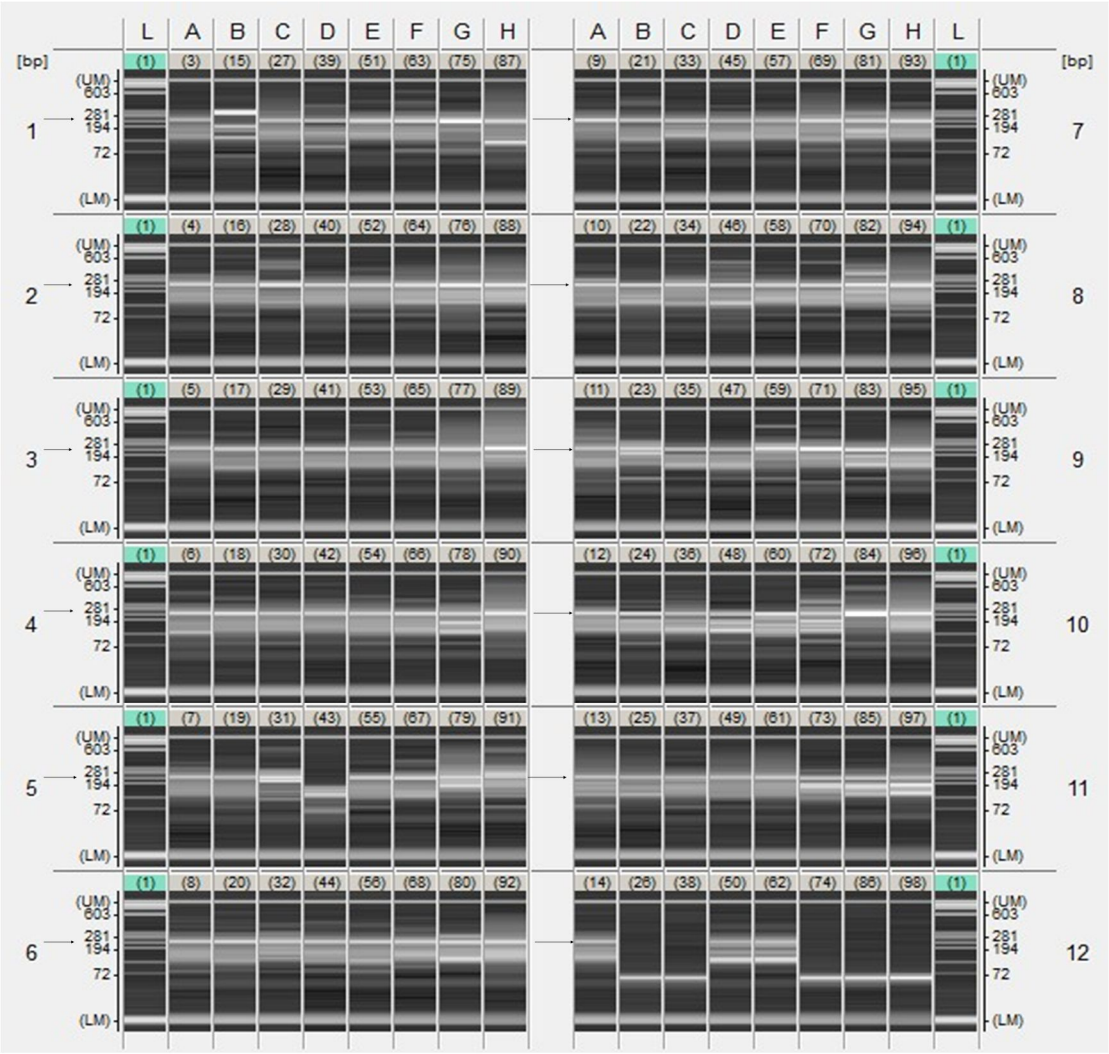


Figure S5. **(a)** Pyruvate Kinase activity, **(b)** Hexokinase activity and **(c)** ratio of Pyruvate Kinase and Hexokinase enzymatic activities in RBCs from control mice and mice transplanted with transduced cells. RBCs were purified from blood samples through a cellulose column to avoid leukocyte PK activity contamination and subjected to enzyme activity evaluation. Black bars, healthy mice (n=2); white bars, mice transplanted with cells transduced with the EGFP expressing vector (n=3); scratched bars, mice transplanted with cells transduced with the coRPK expressing vector (n=3). Checkered bars represent values from a healthy volunteer (n=1). Data represent the average ± SEM of each group.



<u>Figure S6</u>. Gel image of LAM-PCR products generated with Tsp509I enzyme for samples harvested from all mice at different time points and tissues as Fig. S8 shows. Vector integration sites were identified by LAM-PCR amplification of 3'vector LTR-genome junctions. A MultiNA automated system was used, generating a pattern characterized by several bands. Vector backbone derived Tsp509I internal control band (IC) is indicated by an arrow.

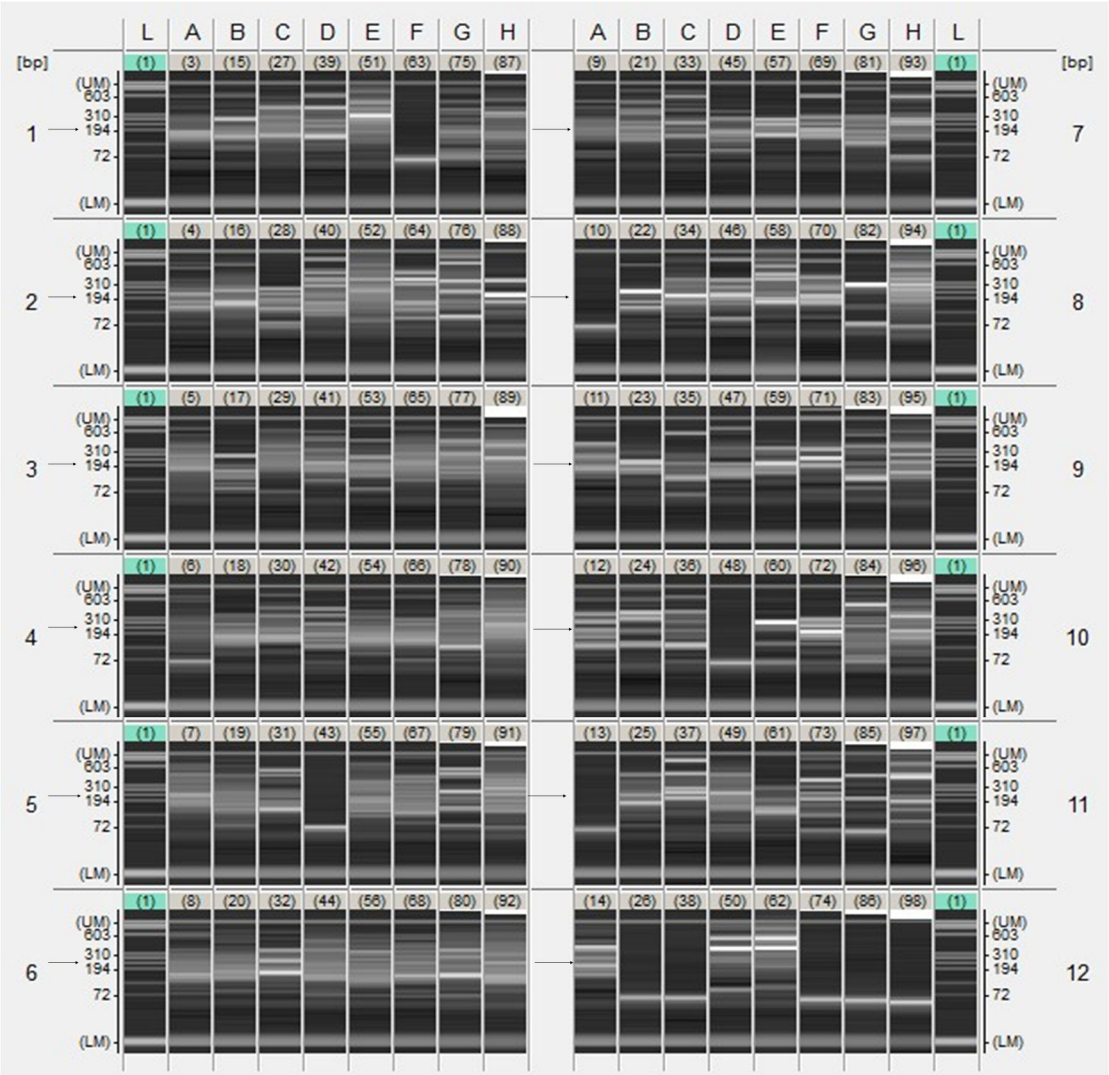


Figure S7. Gel image of LAM-PCR products generated with HpyCH4IV5 enzyme for samples harvested from all mice at different time points and tissues as Fig. S8 shows. Vector integration sites were identified by LAM-PCR amplification of 3' vector LTR-genome junctions. A MultiNA automated system was used, generating a pattern characterized by several bands. Vector backbone derived HpyCH4IV5 internal control band (IC) is indicated by an arrow.

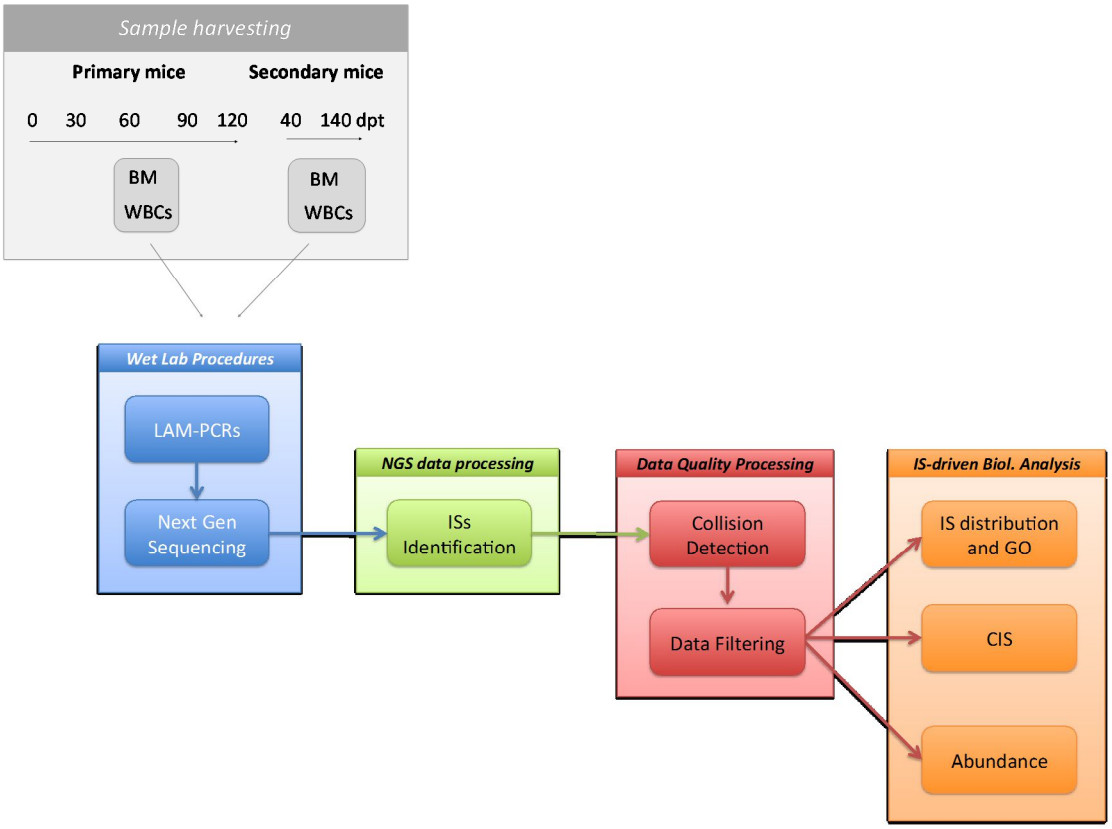


Figure S8. General scheme of the analysis of integration site mapping performed in mice transplanted with genetically modified hematopoietic progenitors. Bone marrow and white blood cell samples from transplanted mice belonging to two independent experiments (**Table S2**) and harvested at different time-points after transplant were analyzed as described in supplementary methods following the showed pipeline.

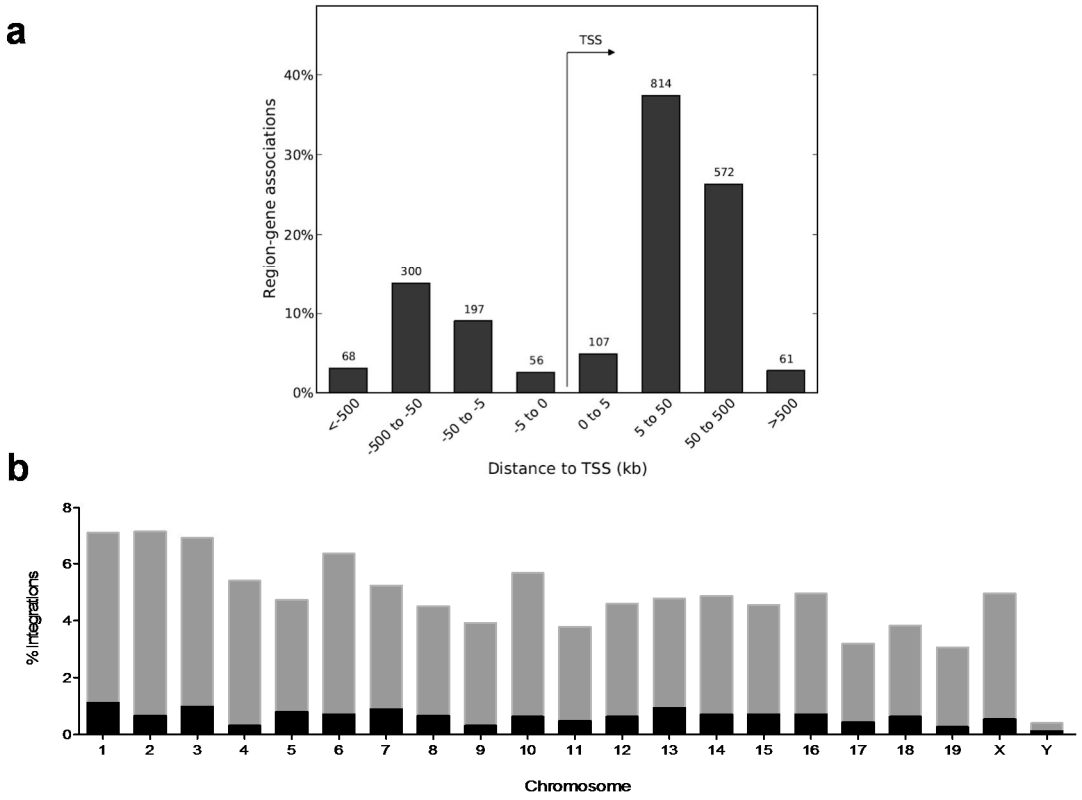


Figure S9. Distribution of LV integrations along the genome of transplanted mice. (a) Integration site (IS) frequency distribution around Transcription Start Site (TSS) of the nearest RefSeq gene, spanning 500 Kb upstream and downstream the TSS. Numbers on the top are the number of IS detected for all samples and time-points. **(b)** Chromosomal distribution of LV integration sites in transplanted mice expressing the EGFP transgene (black bars) or the coRPK therapeutic transgene (grey bars), showing no skewing towards any particular chromosome.

			Primary recicipients				Secondary Recipients			
	Non Redundant IS		99	168	97	19	113	35	32	27
	Total Sequence Reads		14102	340586	314455	69343	206932	152162	200987	180905
Chr	Integration Locus	RefGene	coRPK11	coRPK12	coRPK13	coRPK14	coRPK2.1	coRPK2.2	coRPK2.3	coRPK2.4
16	52114775	**Cblb**			8,87	49,57	5,29	12,35	0,02	
18	35737739	**Matr3**			4,58		4,79	15,84	14,36	17,41
9	59671229	**Myo9a**			2,39		2,69	3,05	0,52	0,74
8	23308857	**Vps36**			6,41		9,34	21,52		4,97
1	41860552	**4930448I06Rik**			1,77		1,68	13,02	8,18	
10	37953519	**Rfpl4b**			1,39	6,17		1,03	0,39	
1	125478355	**Dpp10**			0,91		1,10	3,18		1,63
2	148695277	**Cst3**			0,44		3,02		0,35	2,64
6	80828813	**Lrrtm4**			0,16		0,29	0,44		0,22
7	4446040	**Ppp1r12c**			0,01		0,01	0,01	0,02	
10	7744574	**Tab2**			3,27		9,12			4,07
5	18735403	**Magi2**	33,36	0,81			0,01			
18	57525115	**Prrc1**	17,40				6,26	2,22		
X	105969090	**Brwd3**	1,95	0,71			0,11			
3	148528413	**Lphn2**	1,76	0,45						0,24
5	13643926	**Sema3a**	1,58				0,05	0,53		
4	24449640	**Mms22l**	0,96	0,35			0,11			
1	107487037	**Pign**				12,04			0,35	
19	27467465	**D19Bwg1357e**				9,75		0,63		
13	73458709	**Ndufs6**				9,58			0,00	
14	86078480	**4930529K09Rik**				0,00				0,00
12	72161794	**Arid4a**			9,44				0,00	
1	109398282	**Serpinb2**			5,61					0,00
17	18006611	**Fpr1**			1,35				0,00	
2	38132209	**Dennd1a**			1,28			0,00		
16	81433433	**Ncam2**			0,04					0,00
6	56736516	**Kbtbd2**		6,40					0,03	
7	79224238	**Mctp2**		3,31						22,26
16	58006735	**Col8a1**		3,16					0,00	
4	85596276	**Adamtsl1**		2,37			0,02			
7	111583985	**Trim30a**		2,29			0,01			
16	33948595	**Itgb5**		1,74			0,01			
13	7150736	**Pfkp**		1,11			0,01			
6	9899432	**Nxph1**		0,82			0,00			
4	66828139	**Tlr4**		0,78			0,01			
19	34647666	**Ifit2**		0,70			0,08			
2	174443138	**Zfp831**		0,54						3,79
19	23412889	**Mamdc2**		0,41			0,00			
14	96290784	**4921530L21Rik**		0,33			0,00			
X	82997642	**Gyk**		0,30			0,00			
6	43431106	**9430018G01Rik**		0,16			0,00			
11	98079745	**Cdk12**		0,10			0,00			
6	13808636	**2610001J05Rik**		0,03			0,00			
X	83019292	**Gyk**	3,15					0,00		
4	144607235	**Vps13d**	1,06							0,11
10	53745139	**Man1a**	0,07							0,00
2	104684227	**Prrg4**	0,04					0,00		

<u>Figure S10</u>. Tracked shared integrations between primary and secondary recipient mice carrying the therapeutic PGK-coRPK LV vector. Integrations detected in either mouse in any organ and at any time are pooled. Secondary recipients received the pooled BM from transplanted mice coRPK11 to 14. The rest of the IS detected were detected or in the primary or in the secondary recipients. Numbers in the boxes show the representativeness in percentage of the corresponding integration in the referred mouse. In addition to ≥ 5% filter applied on integration analysis, all integration with a sequence count < 3 were eliminated.

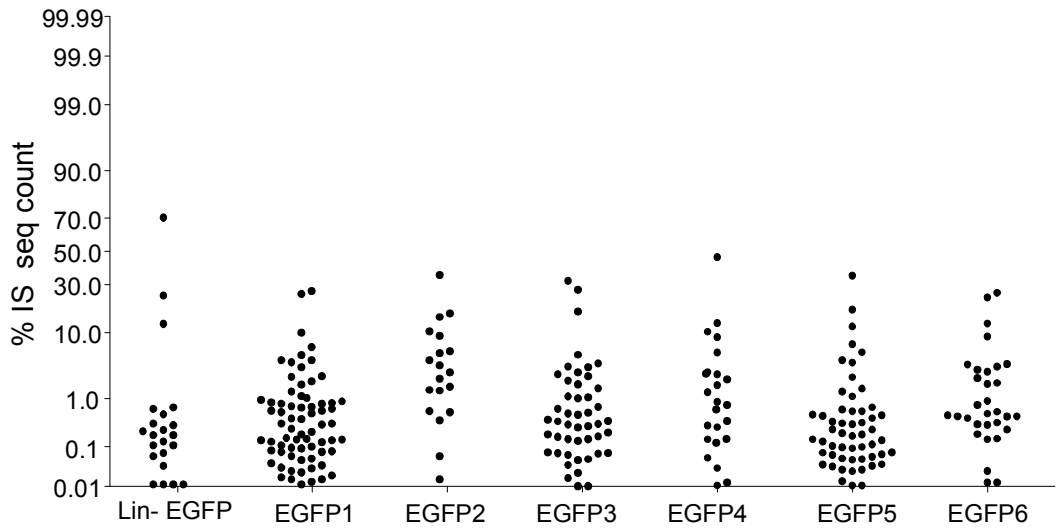


Figure S11. Clonal abundance analysis of EGFP- LV transduced cells. Dots plot representation of clonal abundance of pooled integrations in each mouse in bone marrow**.** The relative percentage (y-axis) for each integration site is relative to the total number of sequences reads obtained in each dataset. Similarly to co-RPK transduced cells (**Fig. 7**), the graph indicates that the vast majority of transplanted mice show a polyclonal pattern of hematopoietic repopulation. IS, Integration site

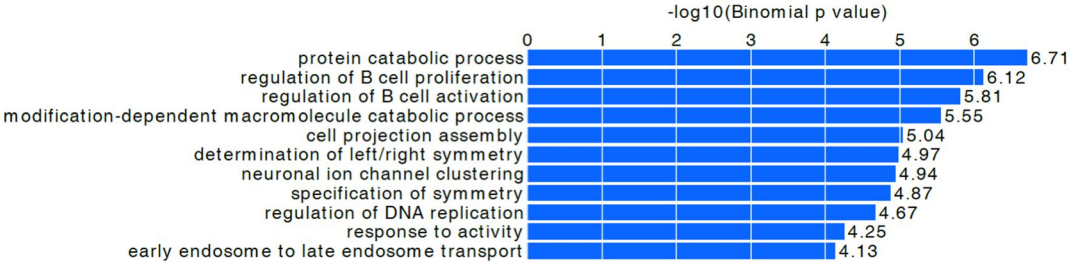


Figure S12. LV genomic integration profile. Gene Ontology (GO) analysis was performed using the GREAT software on samples from transplanted mouse. All integrations retrieved from this study (N=2220) showed overrepresentations of the gene functions indicated on the left part of the figure. To address if the most abundant integrations were enriched on specific gene classes, all integration sites with a relative sequence count >5% of the entire dataset (showed in **Fig. 7**) were selected, showing no GO gene classes overrepresented.

Table S1: Hematological variables recorded 140 days post-transplantation in peripheral blood

Group	HGB (g/dL)	HTC (%)	MCV	MCH (pg)
Healthy (n=5)	14.64 ± 0.99	36.15 ± 2.64	38.20 ± 0.86	15.44 ± 0.21
PKD (n=9)	9.70 ± 0.55	28.91 ± 1.53	51.56 ± 0.50	17.23 ± 0.31
EGFP (n=8)	6.81 ± 0.68	21.32 ± 1.51	49.13 ± 0.72	15.46 ± 0.62
coRPK (n=17)	10.67 ± 0.53*	31.09 ± 1.45*	45.65 ± 0.84	15.35 ± 0.53
2nd coRPK (n=4)	12.40 ± 0.67**	34.11 ± 1.49*	46.25 ± 0.85	16.78 ± 0.23

Data represent the mean ± SEM and were statistically analysed by comparison to EGFP-expressing mice using the Kruskal-Wallis non-parametric test. *$p<0.05$; **$p<0.01$

Table S2: Relevant molecular parameters in mice transplanted with genetically modified cells

Assay	Groups	Vector copy number (VCN/cell)			Transduction % Provirus$^+$ CFUs	Donor chimerism % SRY$^+$ PB cells
		WBC	Total BM	CFU		
1 MOI 1-4	EGFP (n=2)	0.83 ± 0.05	0.42 ± 0.03	0.42 ± 0.00	57.73 ± 12.28	n.d
	coRPK (n=3)	4.76 ± 0.28	3.58 ± 0.34	3.07 ± 0.76	91.08 ± 3.67	n.d
2 MOI 10	EGFP (n=6)	4.56 ± 0.50	4.19 ± 1.29	3.93 ± 0.98	92.32 ± 7.68	61.82 ± 3.61
	coRPK (n=14)	1.65 ± 0.08	0.99 ± 0.13	1.89 ± 0.42	62.06 ± 11.73	63.66 ± 4.45
	2nd coRPK (n=4)	1.44 ± 0.08	n.d	n.d	63.15 ± 0.31[a]	62.89 ± 5.61

Data represent the mean ± SEM, n.d, not determined. [a] estimated transduction percentage obtained by interpolation in the linear regression built from experiment 1 (X axis: VCN/WBC, Y axis: % provirus$^+$ CFUs).

SUPPLEMENTARY METHODS

Vectors and lentiviral supernatant production. LVs were generated as shown in **Fig. S1a**. CoRPK sequence was designed using the GeneArt® software to increase the GC content of the sequence and to prevent cryptic splice sites. Vectors were developed using the pCCL.sin.ppt.hPGK-EGFP-Wpre* construct as backbone, generously provided by Dr. Naldini (HSR-TIGET, San Raffaele Telethon Institute, Milano, Italy). Vector stocks of VSV-G pseudotyped LVs were prepared by 3-plasmid calcium phosphate-mediated transfection in 293T cells (ATCC: CRL-1573, Rockeville, MD, USA), as previously described [Follenzi A, et al. (2000). Nat Genet 25: 217-222]. Titers of infective LVs were determined in HT1080 cells (ATCC: CCL-121) by qPCR as described elsewhere [Charrier S, et al. (2005). Gene Ther 12: 597-606]. Lentiviral stocks of 10^7-10^8 viral particles (vp)/mL titers were routinely obtained.

Purification and transduction of murine HSCs. BM from 8-14 week-old male PKD mice was harvested from the leg bones and lineage negative cells (Lin^-) were purified using the Lin^- Cell Depletion kit (Miltenyi Biotec, Gladbach, Germany), obtaining 70-90% purity. Lin^- cells were pre-stimulated with 100 ng/mL of recombinant human IL-11 (Peprotech EC Ltd., London, UK) and 100 ng/mL of recombinant murine SCF (R&D Systems Inc., Minneapolis, MN) in IMDM-Glutamax medium supplemented with 20% FBS and 0.5% antibiotics (50 U/mL penicillin and 50 µg/mL streptomycin, (Thermo Fisher Scientific, Waltham, MA) for 24h, and then transduced with EGFP or coRPK carrying LVs in two cycles of transduction at MOIs of 1-10 vp/cell (**Fig. S1b**). Each transduction was carried out for 24h in the presence of the aforementioned

cytokines on plates previously coated with CH-296 fibronectin fragment (2μg/cm^2; Retronectin, TakaraShuzo, Otsu, Japan) overnight at 4ºC.

***In vivo* RBC survival.** Transplanted mice carrying the coRPK transgene were injected with three consecutive intravenous injections (12h apart) of Biotin 3-sulfo-N-hydroxysuccinimide ester sodium salt (50 mg/kg) (Sigma Aldrich, Saint Louis, MO). Twelve hours after the last injection, tail vein blood was harvested and labelled with 2 μg/mL of anti-mouse Ter119-PE (BD Bioscience, San Jose, CA) and streptavidin-FITC (50 μg/mL, BD Biosciences, San Jose, CA) for 30 min at 4ºC. Samples were analyzed in an EPICS XL flow cytometer (Beckman Coulter, Brea, CA) every 2-4 days for 40 days after the injection. RBC survival kinetics was measured by the percentage of biotinylated cells within the total RBC population.

CFC Assay. CFC assay was performed in BM and spleen from control and transplanted mice according to manufacturer's procedure from Methocult medium GF M3434 (Stem Cell Technologies, Vancouver, Canada). BM cells were harvested at different time-points after transplant from all groups of mice, and CFUs (clusters of 30 or more cells) were scored 7 days after seeding in a Nikon Diaphot-TMD microscope.

Identification of hematopoietic lineages. PBMCs were obtained from the tail vein of transplanted animals and labelled with a panel of antibodies to detect different hematopoietic cells. Myeloid cells were detected with anti-GR-1 and anti-Mac-1 biotinylated antibodies (BD Bioscience, San Jose, CA, 5 μg/mL), while lymphoid cells were detected using anti-CD3-PE antibody for T-cells, and anti-B220-PE and

anti-B220-PECy5 antibodies for B-cells (BD Bioscience, San Jose, CA, 10 μg/mL), together with SAV-TRC secondary antibody (Invitrogen, Thermo Fisher Scientific, Waltham, MA). Samples were analysed in a BD LSR Fortessa Cytometer (BD Bioscience, San Jose, CA,, USA) adding DAPI (Boehringer, Ingelheim, Germany, 2 μg/mL) to exclude death cells.

Structural and histological studies. Spleens were collected, photographed and weighed on precision scales to determine the presence of splenomegaly. Histological studies were performed on spleen and liver sections obtained following conventional histological methods, and stained with hematoxylin (Gill-2 Haematoxylin, Thermo, Pittsburgh, USA) and eosin (Eosin Alcoholic, Thermo Fisher Scientific, Waltham, MA). Iron deposits were also studied in the spleen by Prussian Blue or Perls' staining (Sigma Aldrich, Saint Louis, MO) following manufacturer's instructions. All sections were examined using an Olympus BX40 light microscope and photographed with an Olympus DP21 camera, with a final magnification of 100x or 200x.

Erythroid differentiation. Flow cytometry analysis of Ter119 and CD71 marker intensities in BM and spleen were used to identify the different erythroid subpopulations as described elsewhere [Socolovsky M, et al. (2001). Blood 98: 3261-3273] using 4 μg/mL of anti-mouse Ter119-PE antibody (BD Bioscience, San Jose, CA,), 10 μg/mL of biotinylated anti-CD71 antibody (BD Bioscience, San Jose, CA,) and streptavidin-tricolor (Invitrogen, Thermo Fisher Scientific, Waltham, MA). Cells were then analyzed in an EPICS XL flow cytometer (Beckman Coulter, Brea, CA) using propidium iodide (IP, 2 μg/mL) to detect live cells.

Provirus quantification. Detection and quantification of integrated provirus per cell was accomplished using complementary primers to the packaging proviral sequence (Ψ) and the mouse Titin housekeeping gene. Total BM and peripheral blood samples were collected periodically, and genomic DNA from nucleated cells was isolated using the DNeasy Blood & Tissue kit (Qiagen, Venlo, Limburg, The Netherlands). Twenty to 50 ng of genomic DNA (gDNA) were amplified by multiplex qPCR using the 7500 Fast Real-Time PCR System (Applied Biosystems, Thermo Fisher Scientific, Waltham, MA) and primers and probes previously described [Charrier S, et al. (2011). Gene Ther 18: 479-487].

Chimerism. Presence of donor cells was quantified by qPCR detecting the Y chromosome SRY gene and the mouse β-Actin housekeeping gene. Primers and probes previously described [Navarro S et al (2006). Mol Ther 14: 525-535] were used and genomic DNA from PB of transplanted mice was amplified using the 7500 Fast Real-Time PCR System (Applied Biosystems, Thermo Fisher Scientific, Waltham, MA). Standard curves were prepared using gDNA extracts from samples containing 0% to 100% of BM cells from male/female mouse mixtures and chimerism was calculated as: % of donor engraftment = $100 \times 2^{(C\beta Act - CtSRY).}$

LAM-PCR procedure. In order to identify vector integration sites, 3' vector LTR-genome junctions were amplified by LAM-PCR following the method published by Schmidt et al. 2007 [Nat Methods 4: 1051-1057]. The starting linear amplification (100 cycles) was performed using biotinylated LTR specific primers and up to 100 ng of gDNA as template. Linear amplification products were purified using streptavidin magnetic beads and followed by complementary strand synthesis, parallel digestion

with 2 different restriction enzymes (Tsp509I and HpyCH4IV) and two ligation reactions using linker cassettes complementary to the ends left by the enzyme's cut. The fragments generated were amplified by two additional exponential PCR steps. LAM-PCR products were separated and quantified by gel electrophoresis on a MultiNA automated system (Shimadzu).

Setup of LAM-PCR products for Illumina MiSeq sequencing. Following the method published by Parazynski et al [Paruzynski A, et al. (2010). Nat Protoc 5: 1379-1395], 40 ng of the second exponential PCR products generated by Tsp509I and HpyCH4IV enzymes were re-amplified using fusion primers containing specific sequences that allow paired end sequencing on an Illumina MiSeq sequencer. LAM-PCR samples were adapted for 454-pyrusequencing by fusion PCR to add the Roche 454 GS-FLX adaptors: adaptor A plus an 8-nucleotide barcode was added to the LTR end of the LAM-PCR amplicon; adaptor B was added to the linker cassette side. In 5'-3' orientation the final amplicon was composed as follow: adaptor A, barcode, LTR sequence, unknown genome sequence, linker cassette sequence and primer B. Purified fusion primer PCR products were run and quantified on a MultiNA automated electrophoresis system, and pooled together in order to obtain a final equimolar library of 10 nM. The final library was then re-quantified using a KAPA Library Quantification Kit for Illumina Sequencing Platform (Kapa Biosystems, Wilmington, MA) on a Viia7 real-time PCR system (Applied Biosystems, Thermo Fisher Scientific, Waltham, MA), obtaining an estimated concentration of 16.35 nM. Finally, libraries were sequenced using the Illumina MiSeq Reagent Kit.

Bioinformatics analysis. To extract vector integration sites (IS) from a high-throughput sequencing platform, both Roche 454 and Illumina MiSeq/HiSeq, a pipeline taking in input the row data (typically in FastQ file format) was designed, providing the list of reliable IS and the nearest gene. Superior level analyses for clonal abundance quantification and gene ontology enrichment were performed using Excel, GraphPad Prism (TM) and available online tools.

NGS data processing and pipeline usage. The step of NGS data processing deals with the management of high-throughput data from Illumina MiSeq sequencing platforms and aims at identify IS in which all valid sequence reads are aligned to the reference genome. Data processing comprises two main activities:

1. **Data quality inspection and analysis**, in which lentiviral vector sequences and other contaminants are trimmed.
2. **Integration site identification**, in which all valid sequence reads are aligned to the genome of reference and valid ISs are retrieved.

Data quality analysis. In order to identify IS from Illumina MiSeq raw data a bioinformatics pipeline was developed. Standard LAM-PCR products contain a LTR sequence, a flanking human genomic sequence and a linker cassette (LC) sequence. The 459 technology allowed retrieval of LAM-PCR sequences with length ranging from 10bp to 900 bp. Similar results were retrieved from Illumina MiSeq paired-ends reads. These length boundaries are important parameters to consider in the quality analysis process since they affect both, the subsequent alignment procedure and the algorithm of the vector components identification. Sequences too short to be correctly aligned to the reference gene were discarded, as well as those exceeding the

maximum size reachable with NSG technology to avoid missing part or all of the LC sequence. Once the pipeline ends for each pool, all integration sites were collected both in files (archived in the TIGET network attached file storage -NAS-) and in the internal database, and maintained in a storage server that keeps track of the modified copies.

Integration site identification. To identify unique integration sites and extract the excel file with all IS in rows and each sample in columns (IS matrix) with the closest gene annotations, we run the following steps:

1. Creation of the IS matrix using the program called *create_matrix*, enabling the collision detection inter projects. This program will produce a tab-separated file (TSV).
2. Annotation of the IS matrix file using the program *annotate_bed*, that will be called as follows for each pool using the input TSV file: *awk '{print "chr"$1"\t"$2"\t"$2}' TSV_FILE | tail -n +2 > TSV_FILE.bed; annotate_bed -a /opt/genome/mouse/mm9/annotation/mm9.refGene.TIGET.gtf -b TSV_FILE.bed -o TSV_FILE.annotated.bed;*
3. Import both annotation and matrix file into a new Excel worksheet, here on called XLS.

Collision detection. In order to obtain a reliable dataset of ISs from each transplanted mouse, we filtered data from potential contaminations/collisions and from false positives based on sequence counts. An additional step of data normalization was required to combine integration sites resulting from different experiments.

The term “collision” is used to identify the presence of identical IS in independent samples. In our experimental setting, the integration of vector in the very same genomic position in different cells is a very low probability event. Thus, the detection of identical ISs in independent samples likely derives from contamination, which may occur at different stages of wet laboratory procedures (sample purification, DNA extraction, LAM-PCRs and sequencing). Although our working pipeline is designed to minimize the occurrence of inter-samples contacts, the high-throughput analysis of ISs intrinsically carries a certain degree of background contamination. Identification of the extent of contamination between samples is crucial also because the retrieval of the same IS in different samples obtained from the same mouse is used in subsequent steps to make inference on biological properties of the vector-marked hematopoietic cells (i.e. multi-lineage potential and sustained clonogenic activity). Thus, we must be able to distinguish the actual occurrence of the same IS in different samples (from the same mouse) from a contamination/collision. To address these issues, we assessed the extent of shared IS among samples derived from different test items and mice as a way to measure the extent of collision in our analyses and then design rules to discard from each mouse’s data set those IS that can be ascribed to collision and minimize the likelihood of scoring false positive when looking for shared IS between samples from the same mouse. We designed a collision detection process allowing the validation of each integration locus. The overall result should be that, given the set I of integration loci, in case of classification of an integration locus i in I as collision, i is discarded from I. We applied collision detection process between 3 independent transplantation groups:

1. **coPKR170s**: mice from assay 1 euthanized at 170 days after transplant with Lin^- cells transduced with the coRPK expressing LV vector (coRPK 1-3).

2. **EGFP**: mice from assay 2 transplanted with Lin$^-$ cells carrying the EGFP expressing LV vector (EGFP 1-6).
3. **coPKR-TC**: mice from assay 2 transplanted with Lin$^-$ cells transduced with the coPKR expressing LV vector (coRPK 1-14), whose blood and BM was analysed at different time-points, including secondary recipients transplanted with a pooled BM from a sub-group of primary transplanted mice (coRPK 11-14).

Each identical IS has different sequence reads (sequence count) among the different mice. Sequence counts can be used to determine whether samples from one mouse contaminated the other mice' samples based on the abundance criterion. In our rationale, an integration found in two mice will be assigned to the mouse that shows the highest abundance, while in the other mouse it will be considered as a contaminant. Therefore, we could identify a threshold of differential sequence count that allows assigning a given collision to a mouse and removing from the others. We retrieved the threshold value from our data obtaining a value of 10, meaning that for each IS, among all TI, if an IS has got an abundance value (percentage sequence count ratio) 10 times lower than the highest abundance value (percentage sequence count) of the other TIs, then it is discarded from the current TI. We applied these rules both among the TI and the selected groups, and Excel files were used to compute the collisions detection by applying the following rules (here detailed for TI filtering but that are extended to groups filtering as well):

1. Isolating each TI, group all samples of the same TI together by summing the sequence counts.
2. For the three TIs obtained, for each IS, compute the percentage ratio of the IS sequence count versus the overall sum of reads for the TI.

3. Then, applying the following rule to compute the decision step with the threshold of 10 that allowed to assigning each IS to a reliable TI.

Once an IS was detected to remove, reads of that IS were removed from the group so that it will no longer assigned also to that group. The filter described above was applied between the mice transplanted with different *ex-vivo* transduced cell populations (one cohort of EGFP expressing mice from assay 2, and two cohorts of coPKR mice belonging to two independent transplantation experiments). Moreover for the coPKR-TC group (assay 2), the filter method mentioned above was modified in two ways:

a) To better highlight the sharing of integrations between time-points in the context of the **clonal abundance analysis** we added the following rule: if one integration is shared between one or more mice then the integration will be kept for all time points even if their sequence count is less than the 10% of the maximum sequence count among mice.
b) For **lineage tracking** relationships we applied a more stringent filter by eliminating the IS with a sequence count lower than 3 and the 10% sequence count filter for sharing between time-points. Meaning that an integration shared between two time points will be kept or discarded only if is more or less than the other respectively.

Gene ontology analysis. All gene ontology analyses were made using the GREAT online software (http://bejerano.stanford.edu/great/public/html/). The web page allow to upload the genomic coordinates of the integrations of each dataset and calculates the enrichment levels in the tested dataset by correlating positional information (based on the binomial distribution analysis for p-value calculations) and annotated function

of the genes nearest to the integration sites (based on the hypergeometric distribution analysis for p-value calculations) [Groeschel S, et al. (2011). J Inherit Metab Dis 34: 1095-1102]. Biological processes and molecular functions of the Gene ontology database were chosen for enrichment analysis. Only the gene classes with a false discovery rate <0.05 for both statistical analyses were considered (**Fig. S12**).

Data storage. All data, both row data and results, are stored in TIGET network attached file storage (NAS) in the root folder, in which all alignments from the pipeline are available, as well as the abundance matrix and plots. NAS storage is secured by authentication and authorization policies, and was built on a reliable and scalable infrastructure using redundancy array of disks RAID 5, and it is under backup on our CrashPlan software registered in TIGET.