This file includes:

Materials and Methods

Supplementary Online Material (SOM) with Figures SOM1 to SOM25 and Tables SOM1 to 19

Supplementary Figure legends and Figures (S1 to S21)

Supplementary Tables S1 to S7

References (*56-61*)

Materials and Methods

Clinical protocol and patients

A phase I/II clinical trial of hematopoietic stem cell GT for the Wiskott-Aldrich Syndrome was initiated on April 2010, after authorization by Istituto Superiore di Sanità on March 15th 2010 and by San Raffaele Hospital Etics Committed on April 2nd 2010 (first approval on December 3rd 2009) (Eudract # 2009-017346-32) (Clinical trials.gov: NCT01515462). The study promoter is HSR-TIGET, San Raffaele Hospital, Italy and the financial sponsor of the study is Telethon Foundation. The medicinal product received Orphan Drug Designation (ODD) (EU/3/12/998) by FDA (ODD#10-3043) and by the European Medicines Agency (EMA) for the treatment of WAS.

Male children with Wiskott-Aldrich syndrome suffering from a severe clinical condition (Zhu clinical score ≥ 3) or severe WAS phenotype (severe WAS mutation/absent WAS expression), without a suitable matched donor for allogeneic transplant or ineligible for HSCT because of age >5 years or clinical features, were eligible for the study. Details on clinical study enrollment criteria, endpoints and study plan, can be found at ClinicalTrials.gov # NCT01515462. The parents of all subjects provided written informed consent for experimental treatment.

Biological samples were obtained from WAS patients, healthy children and adults as controls, with approval of the San Raffaele Scientific Institute's Ethics Committee and consent from parents or subjects.

Lentiviral vector

The lentiviral vector used in this clinical trial is a 3rd generation self-inactivating (SIN) vector derived from HIV, named pCCLsin.cPPT.hw1.6.hWAS.WPREmut (abbreviated as LV-w1.6W), and it has been previously described (5, 31). This is a pseudotype vector made by a core of HIV-1 structural proteins and enzymes, the envelope of the Vesicular Stomatitis Virus (VSV) and a genome containing HIV-1 cis-acting sequences, no viral genes and one expression cassette for the WAS transgene. In the cassette, the endogenous 1.6 KB promoter of WAS controls the expression of WAS cDNA. The three vector components (core, envelope and genome) are transiently expressed in vector producer cells by four different constructs: two core packaging constructs, the envelope construct and the transfer vector construct. Only the vector construct is transferred and integrated into the target cells.

Clinical lots of vectors were produced under GMP conditions by MolMed S.p.a. (Milan, Italy), using a large scale process based on transient quadritransfection in 293T cells followed by purification through endonuclease treatment, anion exchange chromatography with gradient elution, gel filtration, resuspension in serum free media, 0.22 mm filtration and aseptic filling (Figure S1) (31). Each lot was characterized in terms of titer, potency, purity and safety aspects (Table S1 and S2).

$CD34^+$ cell gene transfer

Patients' $CD34^+$ cell manipulation and transduction were performed at MolMed S.p.A., a certified Good Manufacturing Practice (GMP) facility. Flow chart of transduced CD34+ cell manufacturing is shown in Figure S2. Patient's BM was diluted, stratified with Lymphoprep (Axis Shield, Oslo, Norway) and centrifuged in order to collect

mononuclear cells (MNC). $CD34^+$ cell positive selection from BM MNC was performed using immunomagnetic beads (CliniMACS, Miltenyi Biotec, Bergisch-Gladbach, Germany) and an immunomagnetic enrichment device. For peripheral blood leukapheresis $CD34^+$ cells were processed using the CliniMACS device. Phenotypic expression of surface molecules (CD34) was determined by direct labeling using standard methodology. Two independent lots of transduced cells were prepared for Pt1 (one for BM and one for MPB). The number of $CD34^+$ cells collected from the BM and used for transduction ranged from 4.4 to 14.5 x 10^6/kg (Table S3). MPB derived $CD34^+$ cells in Pt1 were 6.5 x 10^6/kg after thawing on day -3 (Table S3). Purified $CD34^+$ cells were seeded on VueLife bags (American Fluoroseal Corp., Gaithersburg, MD, USA) at 1 x 10^6 cells/ml in serum-free CellGro SCGM Medium (Cell Genix Technologies, Freiburg, Germany) in the presence of cells culture grade Stem Cell Factor (SCF) 300 ng/ml (Amgen Inc., Thousand Oaks, CA, USA), FLT3-L 300 ng/ml, Thrombopoietin (TPO) 100 ng/ml and IL-3 60 ng/ml (all from Cell Genix Technologies). Following 24 hour of pre-stimulation, cells were transduced with LV-1.6W at 100 multiplicity of infection (MOI) for 2 hits of transduction with a wash period of 12-14 hours between the vector exposures. Transduced cells were tested for immune phenotype, sterility, endotoxin, mycoplasma, large T antigen, E1A DNA, Large T antigen DNA, clonogenic content and transduction efficiency, replication competent lentivirus (Table S4). Cells were resuspended in saline and transferred in a syringe for infusion. An aliquot of cells was cultured in IMDM, 10% FBS (Cambrex, East Rutherford, NJ, USA) with the same cytokines at 20 ng/ml concentration and harvested after 15 days to perform proviral

integration evaluation by qPCR, WASp expression measurement, and integration profile analysis.

Rationale for conditioning regimen

In contrast to SCID, in which severe differentiation impairments of hematopoietic precursors convey space and a selective advantage for engrafted progenitors with the potential for normal differentiation, the differentiation of haematopoietic cells in WAS is not blocked since all mature subsets are present in the blood of patients, although at reduced levels. Therefore, allogeneic transplant for WAS patients requires full myeloablative conditioning before transplantation, which is usually performed with a combination of high dose chemoteraeuptic agents (Busulfan + Cyclophosphamide + ATG, Fludarabine + Melphalan + ATG/Aletuzemab).

The same principle applies to autologous transplant for gene therapy, but lower intensity of chemotherapy can be envisaged allowing adequate chimerism of gene corrected cells, while reducing toxicity associated with such strong regimens (mucositis, infection, aplasia). Our conditioning regimen is based on administration of iv Busulfan as the main HSPC cell depleting agent, at a dose which is approximately 50% of the one employed in meyloablative alloogneic transplantation. In addition to Busulfan we included Fludarabine and anti-CD20 monoclonal antibodies to favour gene corrected stem and progenitor cells engraftment and further deplete non corrected lymphoid cells. Indeed, the studies on WAS patients with mixed chimerism following allogeneic transplantation or with revertant mutation in the T cell compartment indicated that the selective advantage for WASP+ cells takes place mainly through peripheral expansion upon antigen triggering rather than during thymic differentiation (59,60). Thus, Fludarabine was used

to break the homeostasis in the compartment of early lymphoid progenitors in order to boost, at an early stage of differentiation, the expansion of genetically corrected T cell precursors. In addition, through its activity on the naïve T cell compartment, Fludarabine should favor the establishment of a pool of corrected naïve T cells in the periphery. We also introduced in the conditioning anti-CD20 monoclonal antibody (Rituximab), a biological agent that depletes peripheral B cells and has been shown to provide a clinical benefit in several autoimmune disorders. Anti-CD20 served as depleting agent for B cells and particularly of autoreactive cells, thereby facilitating the engraftment and expansion of gene corrected B cells expressing WAS. In addition, anti-CD20 was used as pre-emptive treatment for lympoproliferative disorder due to EBV, which represents a high risk factor for the development of lymphoma in WAS patients.

Patient's treatment

In Pt1, mobilization of peripheral blood stem cells (PBSC) after G-CSF stimulation was performed to collect back up HSC and store an adequate HSC dose for subsequent transduction and reinfusion in case of need. In Pt2 and Pt3, due to their younger age, only a BM back up was performed. A central venous catheter was implanted in all patients.

On day-3 autologous BM were collected for cell transduction under general anesthesia. A reduced intensity-conditioning regimen was then administered before reinfusion of the autologous-engineered HSC. It consisted in low dose iv Busulfan (bodyweight-based and area under the curve (AUC) adjusted-dose) (range: 0.8-1.2 mg/kg/dose), administered consecutively in 8 doses every 6 hours from days –3 to day –1 and iv Fludarabine (30 mg/sqm/day) on days –3 and –2. The actual doses and AUC of Busulfan received by each

patient are summarized in Table S5. A single dose of anti-CD20 monoclonal antibody (Rituximab 375 mg/sqm) was administered on day -22 to deplete potentially autoreactive B cells and prevent early lymphoproliferative disorder. Transduced autologous LV-wWAS $CD34^+$ cells, manufactured as indicated above, were infused i.v. in 20 minutes. Patients received anti-bacterial, anti-fungal, anti-P. Jirovecii and anti-viral prohylaxes according to local standards. The treatment was administered at the Pediatric Immunology and Bone Marrow Transplantation Unit at the San Raffaele Scientific Institute in Milan, Italy and patients were maintained hospitalized in isolation for 40, 70 and 43 days, respectively

Patients' features and clinical course after GT.

The clinical features before treatment and the current clinical conditions of the first 3 patients treated are summarized in Table 1 (see main text). Pt1 was a 5.7 year-old boy lacking an HLA-compatible donor, with clinical history of recurrent, frequent ENT and viral infections, skin petechiae and moderate-severe eczema (Zhu score: 3). His clinical course after treatment was uneventful. The second and third patients (Pt2 and Pt3) were younger and had a Zhu score of 4, due to history of severe infections and more severe eczema, and GI bleeding. Pt2 displayed also a colitis associated with human herpesvirus 6 (HHV-6) and CMV infection as well as persistently elevated inflammatory indexes and vasculitis-like skin manifestations. Pt3 had important feeding problems, with severe GE reflux and food adversion, and was NG tube-fed. Pt2 developed an autoimmune thrombocytopenia, which is frequently observed after allogeneic transplant and was treated with intravenous immunoglobulins and anti-CD20 mAb. The patient experienced

a Gram negative sepsis complicated by a disseminated Intravascular Coagulation (DIC) and ARDS (Acute Respiratory Distress Syndrome), for which he received steroids for about 5.5 months. Pt3 clinical course was characterized by respiratory infections related to aspiration due to the underlying GE reflux, including an episode of interstitial aspiration pneumonia, complicated by ARDS, and requiring steroid treatment for about 2 months. Both patients recovered well from these events without sequelae. No immune suppressive treatment was administered to the patients subsequently. Platelet transfusions were discontinued at 1, 7 and 4.5 months after gene therapy, respectively. None of the patients experienced clinical manifestations of thrombocytopenia. Pt1 discontinued IVIg 6 months after GT and responded to vaccination with recombinant antigens with production of specific antibodies and in vitro proliferative responses to tetanus toxoid.

Laboratory studies

CFU-C assay

Colony forming unit-cells (CFU-C) assay was performed, immediately after transduction, according to the manufacturer's procedure in Methocult medium (Stem Cell Technologies, Vancouver, Canada). At day 14, colonies were scored to determine number and type of colonies, singly picked and analysed by qPCR to evaluate the percentage of transduction.

Flow-cytometric analysis

Surface staining of transduced HSC was performed with anti-CD34 (8G12) and anti-CD45 (HI30) monoclonal antibodies (BD Biosciences, San Jose, CA). Cells from

peripheral blood of patients and healthy donors, purified by standard density gradient technique (Lymphoprep, Axis Shield, Oslo, Norway), were stained for the expression of surface markers CD3 (SK7), CD4 (SK3), CD8 (SK1) CD19 (SJ25C1), CD14 (MΦP9), CD56 (NCAM16.2), CD25 (2A3), CD41 (HIP8) (BD Biosciences, San Jose, CA, USA) and of intracytoplasmatic expression of WASp. Detection of WASp was performed after permeabilization (Cytofix/Cytoperm kit, BD Biosciences, San Jose, CA, USA) by two rabbit anti-human WASp polyclonal antibodies generated against WASp peptides (one of them was a kind gift of Dr. Hans Ochs) and followed by staining with a secondary Alexa 488 or 647 conjugated goat anti-rabbit antibodies. A rabbit anti-goat Alexa 488 or 647 conjugated tertiary antibodies (all from Invitrogen, Carlsbad, CA, USA) were used. Cells were acquired using a FACSCantoII (BD Biosciences, San Jose, CA, USA) and analysed with Flow Jo Software (Tree Star Inc., Ashland, OR, USA).

Western Blot Analysis

Western Blot was performed as previously described (5, 31), using anti-hWASP (H250; Santa Cruz Biotechnologies, Santa Cruz, CA, USA), anti-hGAPDH (Chemicon, Temecula, CA, USA) antibodies, followed by secondary HRP-coupled antibodies (DAKO A/S, Glostrup, Denmark).

Purification of peripheral blood and bone marrow lineages

CD3, CD4, CD8, CD14, CD15, CD19, CD56 cells were purified from MNC from PB and CD3, CD15, CD19, CD34, CD56, CD61, Glycophorin from BM derived MNCs by

positive selection with immunomagnetic beads according to the manufacturer's procedure (Miltenyi Biotec, Bergisch-Gladbach, Germany).

Determination of Vector Copy Number by Real-time PCR

In order to evaluate the number of lentiviral vector copies integrated per genome, a qPCR was performed using specific primer and probes for human Telomerase and lentiviral vector. A reference standard was obtained from serially diluted transduced human T-cell lines carrying 1 copy of integrated LV. Results of integrated vector copies were normalized for the number of evaluated genomes. As negative control, samples of untransduced cells were used. All the reactions were performed according to the manufacturer's instructions and analysed with an ABI PRISM 7900 sequence detection system (Applied Biosystem, Foster City, CA). Primers and probes sequences have been previously described (31).

Vector integration analyses and bioinformatics.

These methods are described extensively in the ADDITIONAL SUPPORTING ON LINE MATERIAL section.

***In vitro* suppression assays.**

Suppression assays were performed as previously described (56). Briefly, $CD4^+CD25^{high}CD127^{-/low}$ nTreg cells and $CD4^+CD25^-$ effector T cells were isolated from PBMCs by FACS sorting. Cells from HDs were used as control. 15×10^3 $CD4^+CD25^-$ effector T cells were stimulated by CD3-depleted APCs and 1 μg/ml of soluble anti-CD3

mAbs (Orthoclone OKT3; Janssen-Cilag). Suppressive activity of nTreg cells was assessed by co-culture of effector T cells with nTreg cells at 1 to1 ratio. Proliferation was evaluated by 16 hours liquid scintillation counting of 3H-thymidine (Amersham Biosciences) incorporation after 96 hours of stimulation. Cpm: counts per minutes.

Analysis of the NK cell immunological synapse

NK cell immunological synapse were analysed as previously described (61). In brief, NK cells were co-cultured with K562 target cells for 30 minutes to permit the formation of conjugates. F-actin and WASp accumulation, as well as perforin and MTOC polarization were measured to determine synapses maturation. The istance of the MTOC from the synapse was measured using the Imaris Software Package (Bitplane, Zurich, Switzerland), or Velocity software package (Improvision-Perkin Elmer, Lexington, MA). Mean distance of the MTOC to the immunological synapse data were compared by using Mann-Whitney Rank Sum Test.

Chromium release cytotoxicity assay

The ability of NK cells to lyse MHC devoid target cells was performed using a standard Chromium release cytotoxicity assay as previously described (57). Briefly, K562 target cells were labelled with 60 μCi of $Na_2(^{51}CrO_4)$ (Perkin Elmer Inc., Waltham, MA, USA) per 10^6 cells for 2hr at 37°C and resuspended at used concentration. Effectors PBMCs were isolated and co-incubated at serial diluted effectors to target (E:T) ratios of 60:1, 30:1, 15:1, 7.5:1, 3.75:1, in duplicates in a 96 well v-bottom plate (Corning Costar, NY, USA) for 4hrs at 37°C in the atmosphere of 5% CO_2. Radioactivity was measured in the

culture supernatants by Canberra Packard Cobra II Auto Gamma Counter (Canberra Packard, Canberra, Canada). Percent specific target cell lysis was calculated as followed: (experimental.cpm-spontaneous.cpm)/(maximal.cpm-spontaneous.cpm)x100. For spontaneous release, targets were cultured with medium alone and for maximum release in 3% (v/v) Triton X-100 (Sigma, St. Louis, MO) in PBS.

T cell-response to anti-CD3

T cell-response to anti-CD3 was performed as previously described (58). Briefly, 0.1 x 10^6 PBMC were seeded into a round-bottom 96-well plate in IMDM medium containing 5% human AB serum (Lonza, Basel, Switzerland) in presence of coated 1 or 10 mg/ml of anti-CD3 (Orthoclone OKT3, Janssen-Cilag). Proliferation was assessed by 3H-thymidine (Amersham Biosciences) incorporation after 72 or 96 hours of stimulation. Data were expressed as the maximal count per minute (cpm) and stimulation index (SI, calculated as cpm/background cpm).

For proliferation assay on untrasformed T cell lines, overnight starved T cells were stained with 2 μM CFSE and 0.1 x 10^6 cells were cultured in a round-bottom 96-well plate coated with different concentration (from 0.01 to 10 mg/ml) of anti-CD3. Proliferation was assessed after 72 hours evaluating CFSE dilution on a FACSCantoII (BD Biosciences, San Jose, CA, USA) and data were analysed with Flow Jo Software (Tree Star Inc., Ashland, OR, USA).

TCR Vbeta repertoire

24 different TCR Vβ specificities were analysed by flow cytometric analysis by I0Test Beta Mark kit (Immunotech, Marseilles, France) according to the manufacturer's procedure. Cells were acquired using a FACSCantoII (BD Biosciences, San Jose, CA, USA) and analyzed with Flow Jo Software (Tree Star Inc., Ashland, OR, USA).

ADDITIONAL SUPPORTING ON LINE MATERIAL (SOM)

Procedures for integration site mapping

The integration site (IS) analysis was performed according to the flowchart shown below in 4 main macro-activities going from "wet" sample processing to bioinformatics. Each macro-activities comprises several dependent activities, represented by rounded boxes, connected by data flow, represented by arrows (Figure SOM1).

1. *Wet Lab Procedures*: LAM-PCRs and Next-Generation Sequencing (NGS).
2. *NGS data processing*: acquiring as input sequencing reads from NGS platforms and getting as output the list of integration sites.
3. *Data quality processing*: performing three sequential activities to improve data quality. Each activity generates the input for distinct processes in step 4.
4. *IS-driven biological analysis*: performing inferences on safety and efficacy of gene therapy as well as on human hematopoiesis.

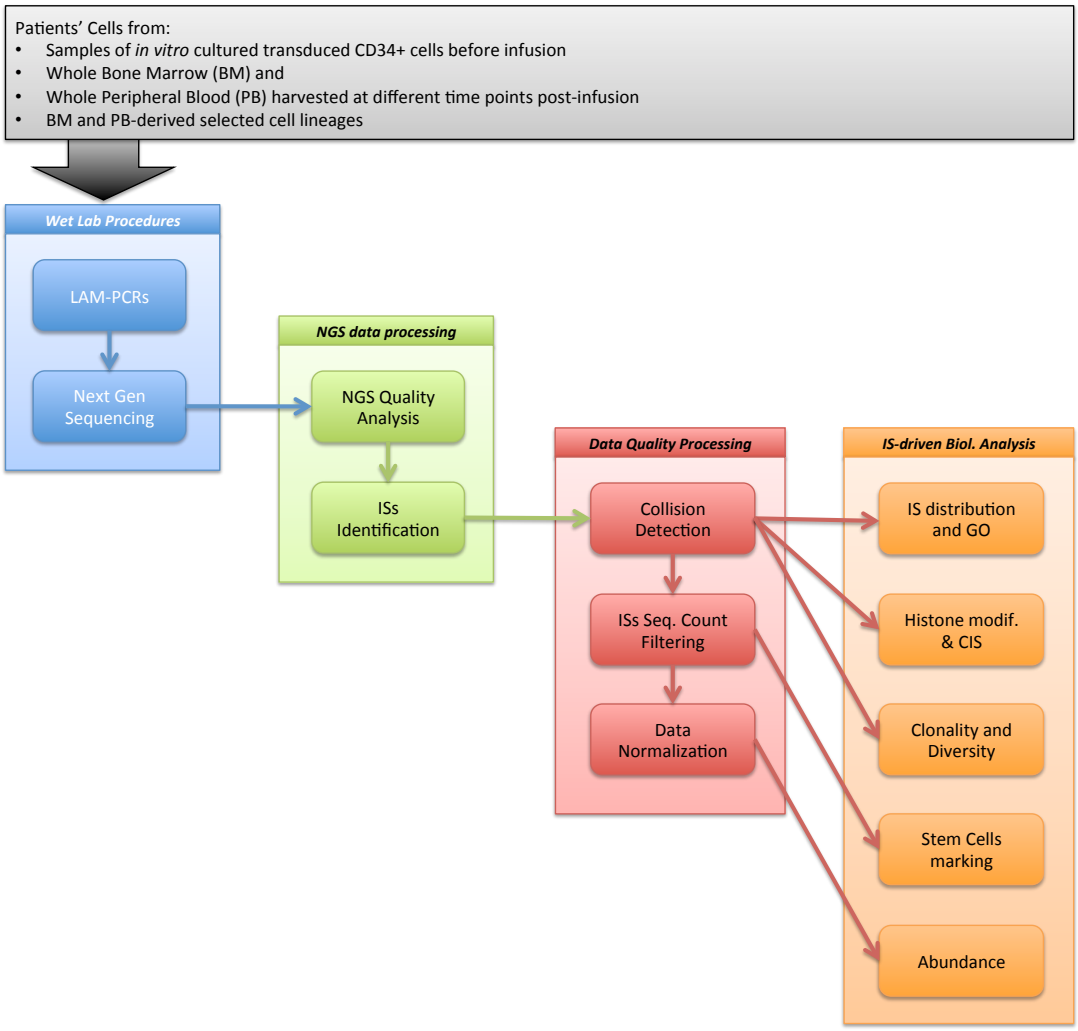


Figure SOM1.

LAM-PCR Procedure

In order to identify vector integration sites 3' vector LTR-genome junctions were amplified by LAM-PCR, according to the method published by Schmidt et al. (42). The starting linear amplification (100 cycles) was done using 5' biotinylated LTR specific primers using 10-100 ng of genomic DNA (gDNA) as template. Linear amplification products were purified using streptavidin magnetic beads and followed by complementary strand synthesis, parallel digestion with 3 different restriction enzymes

(Tsp509 I, HpyCH4 IV and Aci I), and ligation to a linker cassette (LC). The fragments generated were then amplified by two additional exponential PCR steps.

LAM-PCR products were separated by gel electrophoresis on Spreadex high resolution gels (Elchrom Scientific).

454 pyrosequencing of LAM PCR products

Following the method published by Paruzynski et al. (45), we adapted the LAM-PCR samples for 454-pyrosequencing by fusion PCR to add the Roche 454 GS-FLX adaptors and a 6 nucleotide barcode at end containing the LTR sequence (primer A), and the adaptor complementary to the LC (primer B). In 5'–3' orientation, the final amplicon is composed as follows: Primer A, unique barcode of 6 nucleotides, LTR sequence (63 nucleotides), unknown genomic sequence, linker cassette sequence, and Primer B. Each sample was amplified with one of 55 available fusion primers carrying each a different barcode sequence. Fusion-primer PCR products were assembled into libraries avoiding repetition of identical barcodes, and sequenced.

Sample preparation and sequencing on Illumina MiSeq platform

WAS LAM-PCR samples generated by procedure described for 454 amplicon library preparation were quantified with Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen) and checked on a 2100 Bioanalyzer High Sensitivity DNA Chip (Agilent). 250ng of LAM-PCR sample was used for library construction using the TruSeq LT Sample Preparation Kit (Illumina). This involved end repair, adenylation of 3' ends, ligation with indexing adapters, gel purification, PCR enrichment and validation by qPCR with an Illumina

Library Quantification Kit (Kapa BioSystem). Libraries were sequenced using the Illumina MiSeq Reagent Kit (300 cycles) or v2 (500 cycles) for 150 or 250 paired-end reads respectively. In brief, 2nM libraries with different indexes were mixed in equimolar amounts, denatured and diluted to a 10pM final concentration, a genome was pooled at 50% or 70% with these libraries to introduce diversity and optimize the sequencing run performance. All protocols were according to manufacturer's instructions with modifications specified above. To provide run details for each experiment a sample sheet was generated on the Illumina MiSeq instrument. The pair-end runs were initiated for either 2x250 or 2x150 bases of Illunima's sequencing by synthesis technology, including clustering, paired-end preparation, barcode sequencing and analysis. After completion of the run, data were base called, demultiplexed and PhiX reads were filtered. FASTQ format files in Illumina 1.8 format were considered for downstream analysis.

Sequence data processing

The steps of NGS data processing deal with high-throughput data management from Roche 454 pyrosequencing and Illumina MiSeq platforms and it is composed by two main activities:

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

Sequence Quality Analysis

The aim of the NGS quality analysis is to identify the subset of raw reads that are considered valid for further analyses. Our standard LAM-PCR products contain a LTR sequence, a flanking human genomic sequence and a linker cassette (LC) sequence. The 454 technology allowed retrieval of LAM-PCR sequences with length ranging from 10bp to 900bp; similar results were retrieved from Illumina MiSeq paired-ends reads as reported in the plot below showing raw reads counts and the sequence length without LTR and LC segments of all sequences (Illumina paired-ends have been merged at overlapping regions, otherwise we plotted only reads derived from LTR containing sequences).

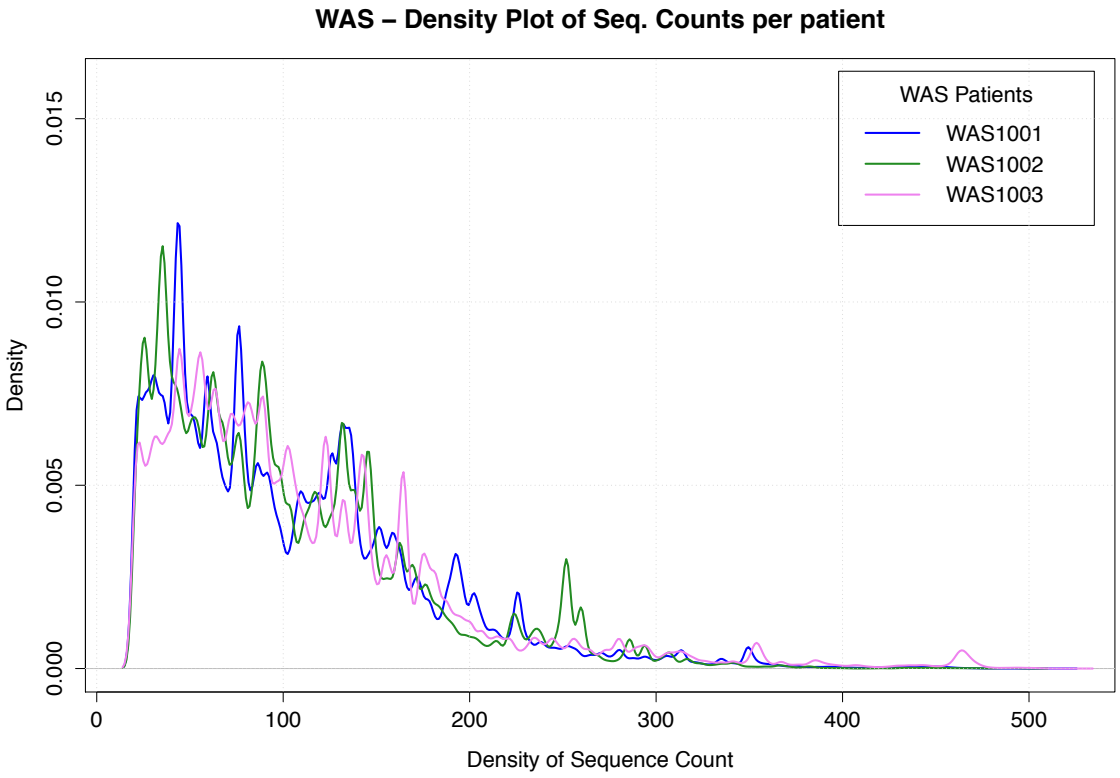


Figure SOM2.

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 vector components identification. If we obtain a sequence too short to be correctly aligned to the reference genome, it will be discarded. If the LAM-PCR product exceeds the maximum size reachable with NGS technologies we could miss part or all of the linker cassette sequence. To address these issues we designed a software tool with customized algorithms and logics for quality analysis in two main steps:

1. LTR sequence identification and trimming: we designed a stringent BLASTN approach (http://blast.ncbi.nlm.nih.gov) for the last 63 nucleotides of the LTR region forcing un-gapped alignment (quality score higher than 99) and perfect match on the last 3 bases. Once the correct LTR fragment is recognized, it is trimmed from the raw read which then flows to the next step. A raw read is discarded if does not have any LTR fragment or the same is below the alignment score quality.
2. LC sequence identification and trimming: reads flowing down from step 1 are checked for the LC presence. A raw read, according to its length, may have or not the LC fragment. We exploited BLASTN algorithm to identify and trim the optional LC presence. The number of output sequences from step 2 is identical to the input from step 1, as step 2 only trims a segment of the input sequence and generates sequences cleaned from both LTR and LC fragments.

Given the initial set of all raw reads ***R***, we run our quality filters obtaining two distinct proper sub-sets of ***R***: ***T***, set of trimmed reads, and ***E***, set of excluded reads. Only reads in ***T*** are considered valid and consistent with our requirements of target analysis and are processed for IS identification.

We implemented the designed software in Python (http://www.python.org) using BioPython (http://biopython.org) libraries for BLASTN analysis and MySQL database (http://www.mysql.com). Software details (under Linux Ubuntu OS, http://www.ubuntu.com):

- Python release 2.6.5
- NCBI BLASTN (version 2.2.18)
- MySQL database release 5.1.41

In Table SOM1 we summarize the results of quality analysis process.

Table SOM1.

Patient	Number of LAM PCRs	Raw reads	Trimmed reads	Discarded reads
1	224	4,338,522	2,771,472	1,567,050
2	239	2,568,207	1,750,253	817,954
3	160	1,798,043	1,399,160	398,883

ISs Identification

To identify *bona fide* ISs it is required to precisely recognize the first bases of the vector-genome junctions using customized and stringent parameters.

Starting from the set of trimmed reads $\boldsymbol{T}$, we designed to map each $\boldsymbol{t}$ in $\boldsymbol{T}$ to the human reference genome (build HG19/GRCh37, Feb. 2009) with NCBI BLASTN tool. To avoid mapping biases, we discarded all reads with length below 20 bases. For each $\boldsymbol{t}$ we collected all alignment output data (mapping score, identity score, e-value, gaps, q-size, span, starting/ending base, etc.) and all mapping positions.

We designed a pipeline with cascading rules to correctly identify each integration site. Each aligned read ***t*** must have:

1. An alignment score greater than 95%: this threshold allows selection for reads with a high identity alignment value.
2. The starting alignment position within the third base: we set this threshold since an integration site could span in a range of +/- 3 bases with respect to aligned reads.

The pipeline evaluates rules and produces from the input set three new subsets:

- ***U***, set of unambiguous (univocally mapped) sequences that comprises all valid integration sites.
- ***A***, set of ambiguous reads with multiple alignments.
- ***N***, set of "no-hit" reads without any positive match to the reference genome.

Elements in U are mapped reads corresponding to so called "redundant" ISs, meaning that each IS may have one or more reads ***u*** in ***U***. To generate a list of unique ISs, we grouped mapped reads based on their alignment starting position in a range of +/- 3 bp. We defined the following rule (called no-redundancy rule): for each read ***u*** in ***U***, list all integration loci, ***L,*** sorted ascending by chromosome and base pair position. We know that $|L| <= |U|$ thus, for each ***l*** in ***L*** we have at least one corresponding mapped read ***u*** in ***U***. If the difference in position between consecutive elements ***l*** in ***L*** (analyzed chromosome by chromosome) is below the threshold of 3bp, then only the first position is the representative one of the integration locus ***i***; otherwise the different positions are considered representative of distinct integration sites. An additional algorithm specifically designed for reads mapped into repetitive elements allowed us to increase the set of redundant sequences.

We implemented a software tool in Python programming language to address each step of the algorithm, exploiting a MySQL database to store data and results. Table SOM2 shows counts of integration sites for each patient.

Table SOM2.

Patient	LAM PCRs	Trimmed reads	Reads of ISs	Identified ISs
1	224	2,771,472	1,095,238	11,530
2	239	1,750,253	755,325	11,158
3	160	1,399,160	579,270	10,576

From the original dataset generated by NGS sequencing, 26-33% of reads could be used as sequences to process for the no-redundancy step. These numbers are different from classical genomic studies in which the amount of discarded reads is usually less than 10%. This difference is due to specific technical aspects of the wet-lab procedures and the custom rules required to identify integration sites. In fact, the majority of reads we discarded (15-50%) were vector sequences amplified as internal controls of LAM-PCR (as reported in Table SOM1). Moreover, as shown in Table SOM2 we collected a set of reads aligning to repetitive elements that were still potential/candidate integration sites. However, since we could not precisely identify unique genomic locations for them, we decided to discard this set of sequences decreasing the fraction of raw reads available for further analysis.

Unique ISs sequences from retroviral WAS GT clinical trial (13) were kindly provided by Prof. Christof Von Kalle and Dr. Manfred Schmidt. In order to obtain comparable data

for the analysis of the two trials, we exploited our pipeline to map these sequences producing the results in Table SOM3.

Table SOM3.

Patient WAS Germany	Original Unique ISs	Mapped Unique ISs
1	5,364	4,209
2	8,742	7,085

Data quality processing

In order to obtain a reliable dataset of ISs from each patient, 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.

Collision detection

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 NGS). 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 patient is used in subsequent steps to make inference on biological properties of the vector-marked hematopoietic cells (i.e. multilineage 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 patient) from a contamination/collision. To address these issues, we assessed the extent of shared IS among samples derived from different patients as a way to measure the extent of collision in our analyses and then design rules to discard from each patient'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 patient (see also later). 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 for data derived from all patients of our two concurrently evaluated and on-going clinical trials of lentiviral vector gene therapy (3 WAS patients, the present study, and 3 Metachromatic Leukodytstrophy MLD patients , for a total of 32.616 IS, from a study reported separately). We analyzed each patient's collisions with respect to the other patients, therefore all counts and filters are patient-based. A summary of shared integration sites for each patient is represented in Table SOM4.

Table SOM4.

Patient	ISs	Collisions Inter-Trials
		WAS + MLD
1	11,530	393
2	11,158	269
3	10,576	239

Each identical IS has different sequence reads (sequence count) among the patients. Sequence counts can be used to determine whether samples from one patient contaminated the other patients' samples. Therefore, we could identify a threshold of differential sequence count that allows assigning a given collision to a patient and removing from the others. We retrieved the threshold value from our data by analyzing each patient independently and then combining all patients' results, as follows:

Given $\boldsymbol{C}$ the set of collisions, each $\boldsymbol{c}$ in $\boldsymbol{C}$ has a sequence count $\boldsymbol{s}$. For each patient, we independently analyzed all collisions: given $\boldsymbol{P}$ the set of patients, $\boldsymbol{p}$ the current patient and $\boldsymbol{p}_{\{n\}}$ all other patients ($\boldsymbol{P}$-$\boldsymbol{p}$), for each patient $\boldsymbol{p}$, for collision $\boldsymbol{c}$ in $\boldsymbol{C}_{\boldsymbol{p}}$ we computed the ratio $\boldsymbol{s}|_{cp}/\boldsymbol{s}|_{cp\{n\}}$, called collision relative frequency (*ColRF*). We then analyzed the distribution of all *ColRF* in $\boldsymbol{C}$ to look for flexes and peaks. Figure SOM3 shows a representative *ColRF* plot, where all *ColRF* values are in log_{10} transformation.

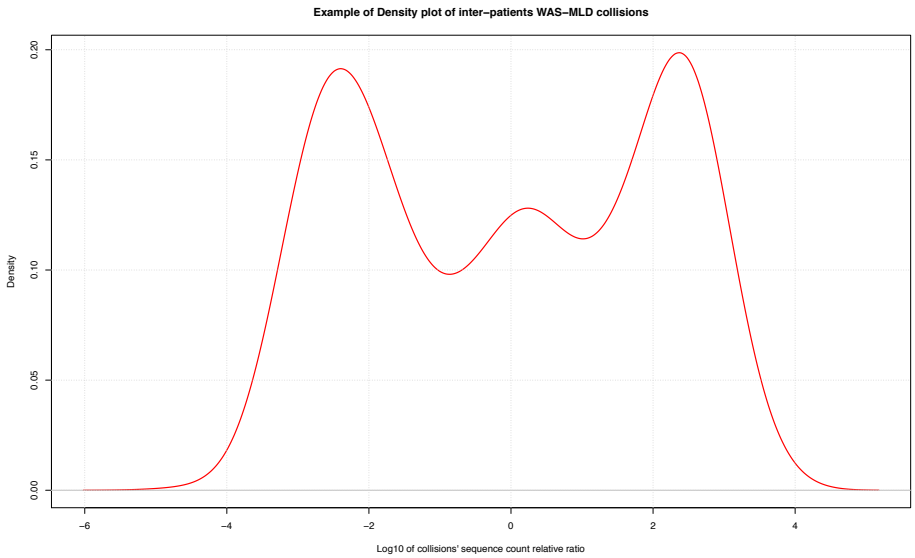


Figure SOM3.

Figure SOM3 is an example of collision relative frequency in 454 sequencing data and shows peaks at around +/-2 and 0. Positive peak at +2 means that all collisions of this area carry sequence counts 20 times higher in the analyzed patient with respect to the others. On the other hand, the collisions under peak at -2 have 20 times lower sequence counts in the same patient as compared to the others. The peak at 0 indicates that all these collisions have identical sequence counts among patients. Table SOM5 shows theoretical use case scenarios for two patients.

Table SOM5

IS id	**Patient A**	**Patient B**	**Ratio A/B**	**Log10 A/B**
1	1	100	0.01	-2.00
2	20	100	0.20	-0.70
3	100	100	1.00	0.00
4	150	100	1.50	0.18
5	250	100	2.50	0.40
6	1100	100	10.01	1.00

Applying these theoretical scenarios to our empirical *ColRF* curve, we interpreted data as decision plot, shown in Figure SOM4. Thus the threshold for contamination identification patient-based is set on value 1, corresponding to 10 fold difference in linear scale.

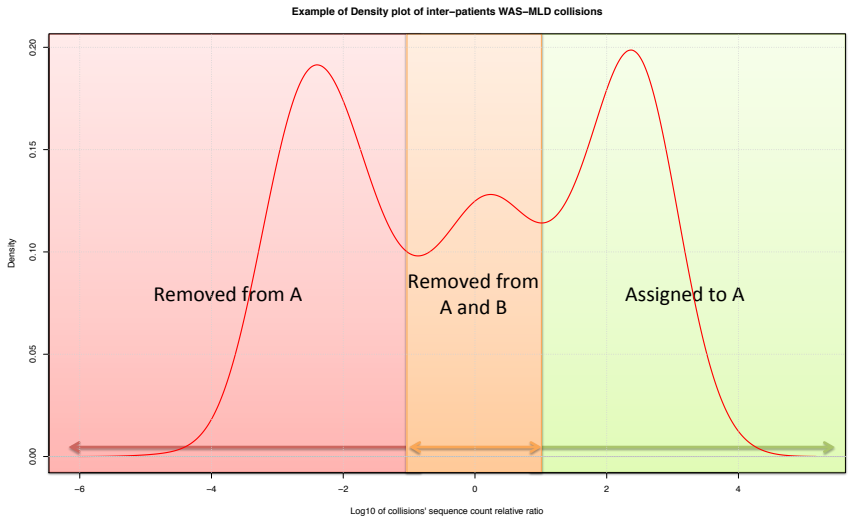


Figure SOM4.

Table SOM6 shows resulting decision fields.

Table SOM6.

IS id	Patient A	Patient B	Ratio A/B	Log10 A/B	Removed from A	Removed from B	Assigned to A	Assigned to B
1	1	100	0.01	-2.00	X			X
2	20	100	0.20	-0.70	X	X		
3	100	100	1.00	0.00	X	X		
4	150	100	1.50	0.18	X	X		
5	250	100	2.50	0.40	X	X		
6	1100	100	10.01	1.00		X	X	

Table SOM7 and Table SOM8 lists integration sites removed from patients 1, 2 and 3 after collision filtering.

Table SOM7.

Patient	ISs	Collisions Inter-Trials	Removed ISs
		WAS + MLD	
1	11,530	393	174
2	11,158	269	151
3	10,576	239	192

Table SOM8.

Patient	ISs	ISs remaining after collisions
1	11,530	11,137
2	11,158	10,889
3	10,576	10,337

In order to understand where potential contaminations occurred among patients and between our trials we further analyzed the sequences that were filtered out by collision detection process (Table SOM9).

Table SOM9.

	Pt.s	Unique ISs Removed from Study	ISs Removed from study	Removed ISs seq. Count	MLD Assigned to pt.			WAS Assigned to pt.		
					1	2	3	1	2	3
WAS	1	466	174	9,078	170	4	9	-	29	7
	2		151	4,794	11	13	9	17	-	68
	3		192	5,074	9	6	4	3	25	-
MLD	1	607	296	5,148	-	78	9	129	3	14
	2		223	6,788	145	-	15	13	11	130
	3		149	3,964	68	258	-	10	10	16

As an example, in Patient 1 out of 11530 ISs we counted 393 collisions of which 174 were removed from the overall study (not re-assigned to other patients) with a cumulative sequence count of 9,078 (1.51% of ISs, representing 0.83% of sequence reads).

Figure SOM5 shows the number of detected collisions in WAS and MLD clinical trials. We removed 466 IS from WAS study: 362 collisions were removed from WAS study only, 104 were removed both from WAS and MLD study, while 503 ISs were removed from MLD study only.

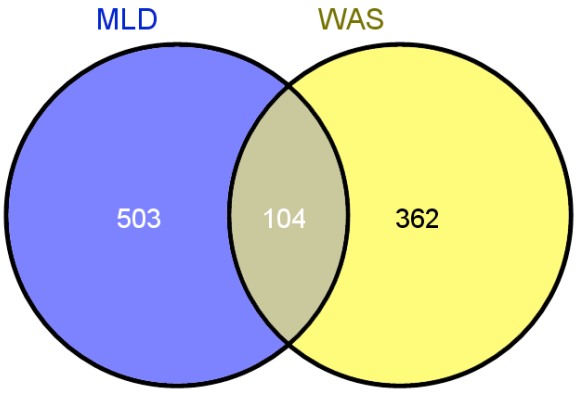


Figure SOM5.

Overall, considering that we retrieved a similar total number of IS from each patient, we did not observe any major prevalence of contaminations derived from a single patient with respect to the others, nor among patients of the same versus different trial. These findings are in line with an expected stochastic occurrence of contamination among samples processed for high throughput sequencing of integration sites and provide an estimate of the extent of occurrence of such cross-contamination between samples in our experimental setting averaging 1.64% of the IS retrieved, with an average sequence count of 0.7% from an individual sample.

The integration datasets cleaned of collisions (available upon request) were used for all subsequent analyses.

Integration sites filtering based on sequence count

As previously described, each IS is associated to a sequence count value corresponding to the number of reads that map into the relative locus. These parameters are important for the validation of ISs (as in the collision detection process) but also for other biological data mining based on ISs. Therefore, in order to generate adequate customized filters for further analysis we performed a detailed study of sequence counts properties in our datasets.

To analyze the distribution of sequence counts values we plotted ISs data cleaned from collisions on a boxplot representation, As an example, box plots of ISs identified in CD34+ cells from each patient over time are shown in Figure SOM6 in which sequence counts data are represented in *log10* scale.

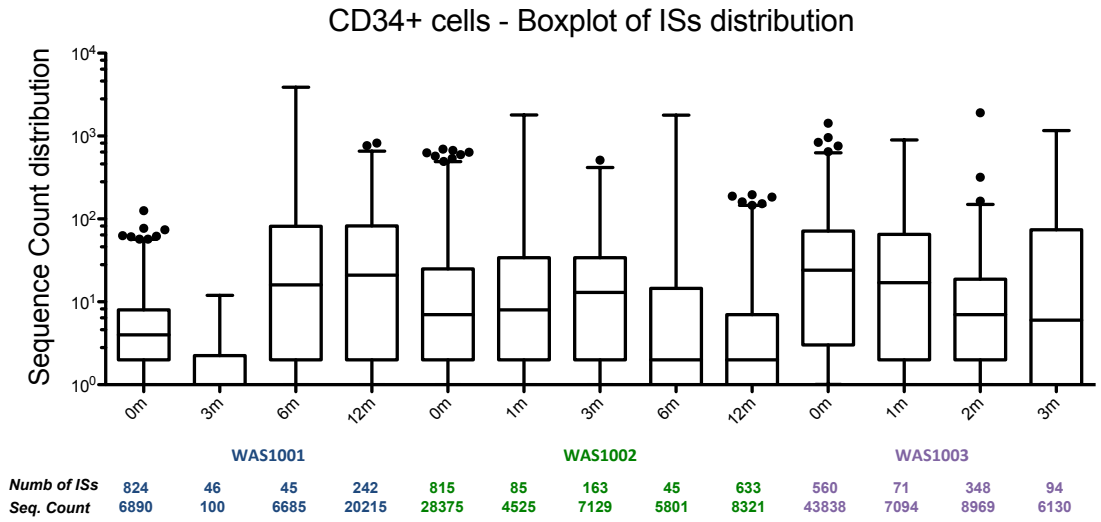


Figure SOM6.

Several boxes on this picture were flattened on the X-axis value 1 suggesting that a relevant portion of integration sites carried only 1 corresponding sequence. Some biological analyses require stringent parameters for ISs validation to minimize sequencing or PCR artifacts. Therefore, for this set of analyses we applied a threshold for the minimum number of reads for each integration site with the goal to minimize the number of false positives. We tested the threshold of sequence count 3 to exclude the ISs represented by very low counts, which carry a higher probability of being affected by potential sequencing errors. When applying this threshold to the boxplot analysis shown in Figure SOM7.

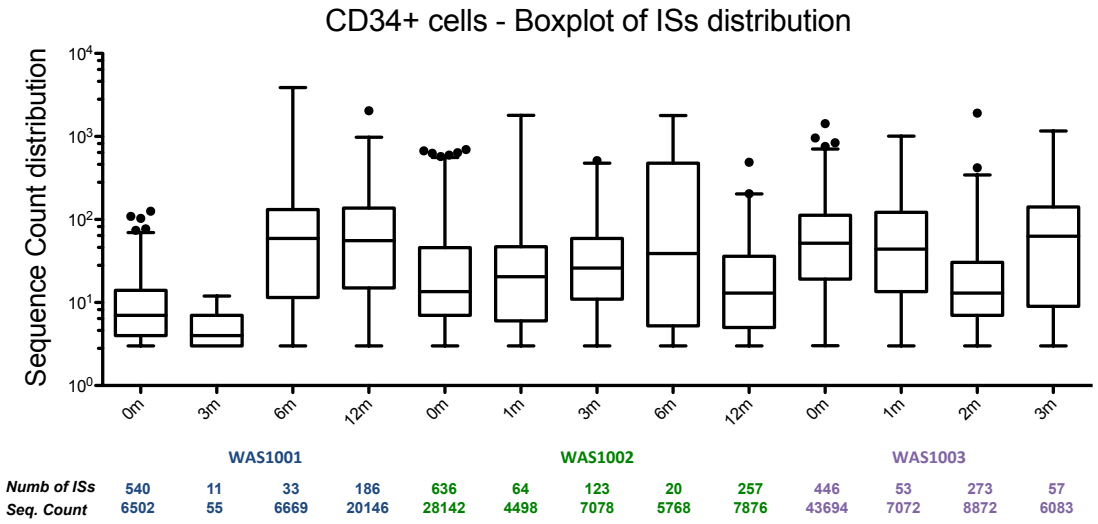


Figure SOM7.

In another representation based on Gaussian kernel density distributions (Figure SOM8), IS sequence counts are shown before and after filtering for the threshold of 3 per IS in Figure SOM7 for CD34 over timepoints.

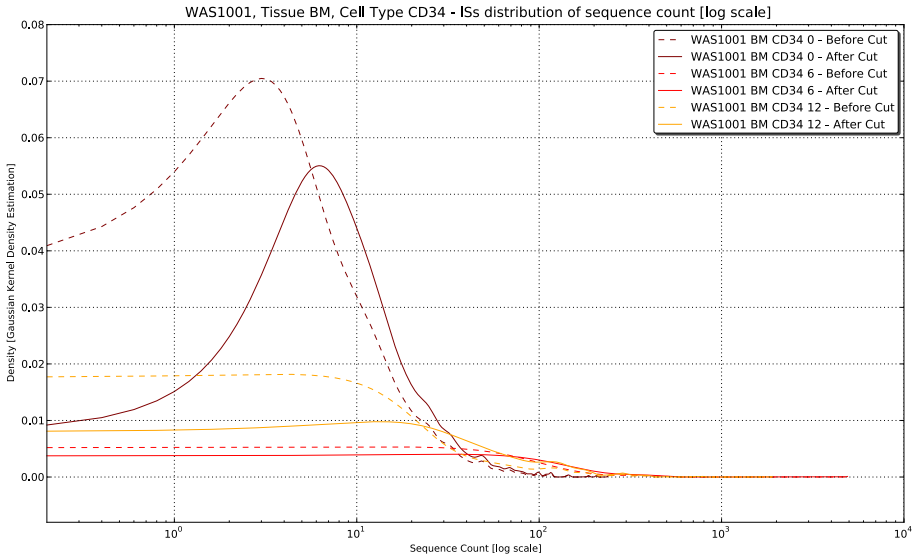


Figure SOM8.

Table SOM10 summarize resulting IS after filtering up to 12 months.

Table SOM10.

Patient	Integration sites	Filtered by Collisions	Filtered by Seq. Count
1	11,530	11,137	8,323
2	11,158	10,889	7,905
3	10,576	10,337	7,836

We applied this filtering procedure to all datasets undergoing analysis of stem cell marking and, after data normalization, of clonal abundance.

- In stem cell marking analysis, ISs with a sequence count lower than the threshold of 3, would have affected the analytical stringency, leading to potential overestimation

of the number of transduced stem cells (see section of Stem cells marking analysis for details).

- In abundance analysis, we estimated that preservation of ISs with a sequence count lower than the threshold of 3 would have had a negative impact on background noise upon data normalization.

Common insertion sites detection

Vector integration frequency along the genome is not homogeneous. Dense clusters of integrations contained in a relatively narrow genomic interval, known as Common Insertion Sites (CIS) have been used as an indicator of ongoing genetic selection and enrichment of cell clones harboring integrations that, by targeting specific genes, have acquired a selective advantage *in vivo*. In hematopoietic cells from patients from the γRetroviral Vector (γRV) –based clinical trials for X-SCID, CGD and WAS, CIS were identified. Among all CIS identified, some, targeting cancer genes such as *LMO2*, *MECOM*, *PRDM16*, *CCND2* and *SETBP1*, were found in leukemic/dysplastic/dominant cell clones from patients' blood.

To investigate the presence of CIS in our study we used:

1. A region-based approach based on sliding windows (49).
2. A method for CIS identification based on a new genome-wide Grubbs test for outliers' analysis.

Region-based CIS identification approach based on sliding windows

Using Abel at al. R package (49) (latest update of August 2012), the function "Cluster" returned the original input list of ISs with additional annotation fields such as: "CIS max

order" that represents the maximum number of integrations contained in each CIS and "Cluster ID" that represents a genomic window in which one or more CIS intervals are clustered.

CIS intervals of a given max CIS order may contain several genes targeted by a number of integrations lower or equal than the CIS order. For example in patient 1 a CIS of max order 33 contained 5 genes, (UNKL, CLCN7, TELO2, TMEM204, IFT140) in which UNKL was targeted by 13 integrations while the remaining 20 integrations were distributed among the other 4 genes (CLCN7 8, TELO2 3, TMEM204 4, IFT140 5).

To represent these data, for each patient we grouped all ISs targeting the same gene (flanking regions up to nearest annotated gene included) and plotted the results in histograms (Figure SOM9-11). The red bars represent the CIS max order value for 100 genes contained within the top ranking CISs. The blue bars represent the number of integration sites within each gene sorted by genomic position within each CIS cluster. Green bars group genes by clusters (under Abel et al. method implementation).

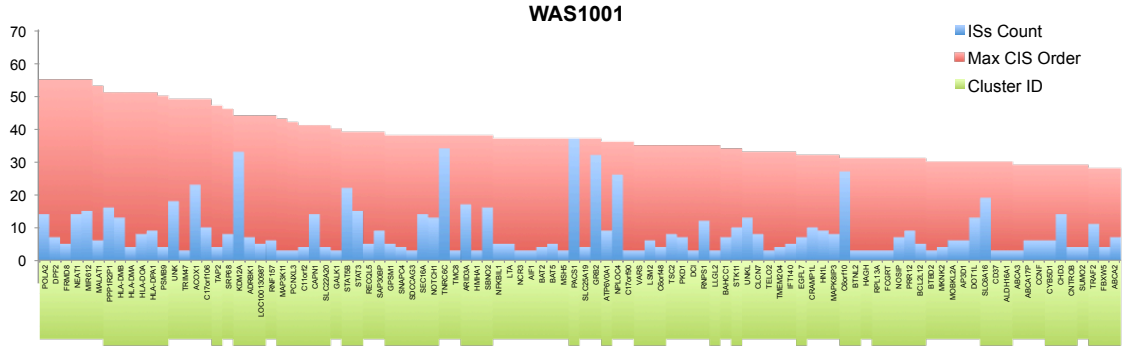


Figure SOM9.

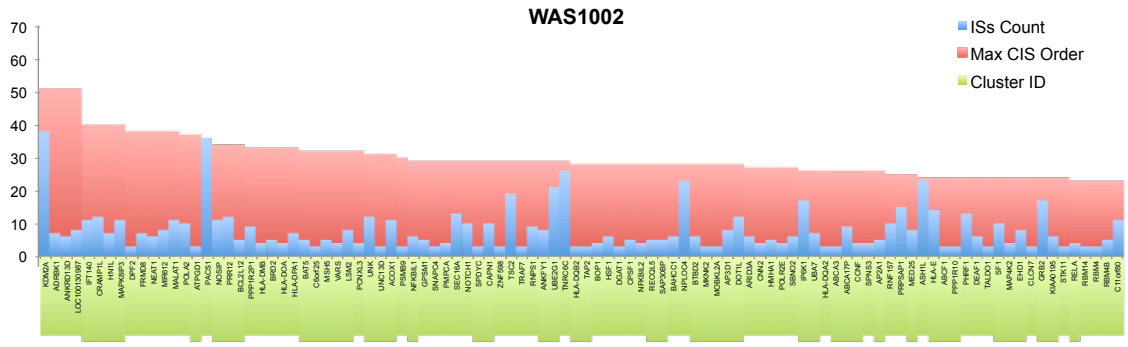


Figure SOM10.

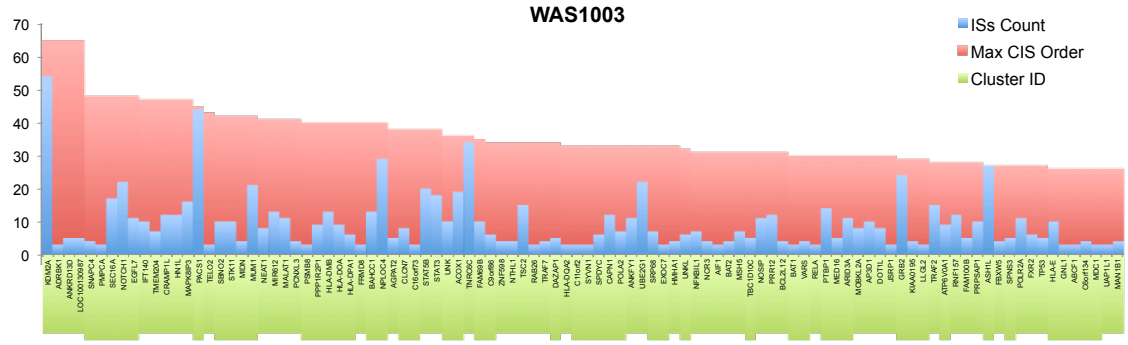


Figure SOM11.

Selection of the most targeted genes within the CIS intervals identified by Abel's method

To better summarize which genes were more targeted within the CIS intervals identified by this method we selected CIS genes of max order 10 and targeted by at least 7 integrations. We then analyzed the number of CIS genes shared among our 3 patients (Table SOM12, Figure SOM12).

Table SOM12.

Patient	Clusters	CIS Genes order >7	Top 20 targeted CIS
1	150	201	PACS1 (37), TNRC6C (34), KDM2A (33), GRB2 (32), RPTOR (28), C6orf10 (27), NPLOC4 (26), TNRC6B (25), ASH1L (23), ACOX1 (23), RERE (22), STAT5B (22), NSD1 (21), FCHSD2 (20), FANCA (20), GPATCH8 (20), IP6K1 (19), SLC6A16 (19), EIF4G3 (18), FNBP1 (18)
2	134	181	KDM2A (38), PACS1 (36), TNRC6C (26), FCHSD2 (24), ASH1L (23), NPLOC4 (23), TNRC6B (22), UBE2G1 (21), RPTOR (21), TTC28 (21), PCDH9 (20), NFATC3 (20), TSC2 (19), RERE (18), C6orf10 (18), SMG6 (18), IP6K1 (17), GRB2 (17), EIF4G3 (16), NSD1 (16)
3	129	183	KDM2A (54), PACS1 (44), TNRC6C (34), NPLOC4 (29), ASH1L (27), C6orf10 (25), EIF4G3 (24), GRB2 (24), NOTCH1 (22), UBE2G1 (22), FBXL20 (22), TNRC6B (22), MUM1 (21), NF1 (20), STAT5B (20), VAV1 (20), PBX3 (19), FCHSD2 (19), ACOX1 (19), NSD1 (18)

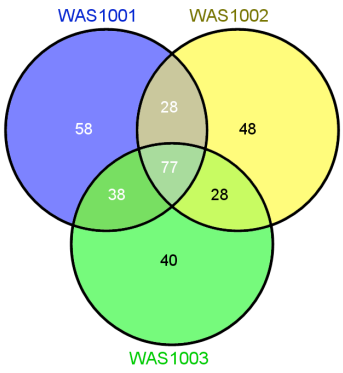


Figure SOM12

CIS Analysis by genome-wide Grubbs Test for outliers

Not all CIS are necessarily the product of genetic selection as several CIS are also found in cells at early time points after transduction, before genetic selection can occur. We

have previously shown that LV CIS in blood cells from HSC gene therapy patients of Adrenoleukodystrophy (27) and in human/mouse hematochimeras are clustered in specific megabase-wide genomic regions heavily targeted by LV integrations (50). On the contrary, the CIS derived from insertional mutagenesis are not clustered in the genome and typically target only one gene within a genomic region, the culprit of the selective advantage/transformation. This observation, together with previous data from other laboratories, suggested that LV CIS might originate as a consequence of intrinsic integration preferences of this vector for specific genomic regions rather than the result of insertional mutagenesis.

To distinguish intrinsic integration hotspots from those resulting from genotoxic selection we applied the Grubbs test for outliers to compare the integration frequency at genes contained in selected megabase wide genomic regions. Differently from other approaches for CIS analysis, our approach is gene-centered and corrects the integration frequency by the size of the targeted gene rather than by user-defined genomic intervals (49). Analysis of the Z-score distributions of the gene integration frequency allows identifying those genes that have been targeted at a significantly higher frequency with respect to others focusing on a given genomic region of interest. Therefore our approach does not assume a random distribution of the integrations along the genome, rather takes into account the average gene integration frequency observed within the selected genomic region, the variance of integration frequency and the number of observations (targeted genes). This approach reflects our previous "regional" Grubbs test for outliers to identify CIS in a genome-wide fashion. Although the statistical principles and the procedures for CIS identification by genome-wide Grubbs test for outliers' analysis are essentially the same

as the previous one, there are also some important differences. Here we compare the integration frequency of each targeted gene to the average frequency of all genes of the dataset without correcting for regional biases. This allows ranking genes by their relative integration frequency with respect to the whole genome and better highlighting the most highly targeted regions. Moreover, also the number of genes analyzed is very different, as in regional analyses the genes may range between 7 and 80, while in our genome wide approach they can be from 500 to thousands. The latter is an important difference as the number of genes analyzed impacts on the significance because the p-value obtained is corrected (multiplied) by the number of genes analyzed. Given the large number of genes evaluated in genome wide analyses, this correction greatly increases the p-values and leads to loss of significance. To avoid the high toll imposed by this type of p-value correction we opted for a multi-step p value correction. Firstly, we generated a list of CIS genes with non-corrected p values <0.05 (raw p-value). The list of CIS genes with a raw p-value<0.05 is useful for the visualization and ranking of the top gene integration frequencies. To reduce the false discovery rate we then corrected the raw p-value by the number of significantly overtargeted genes before correction (raw p-value x Number of genes with a raw p value <0.05). Finally, the genomic regions comprising the CIS genes selected by this more stringent correction were used for regional Grubbs analysis to correct for local biases of integration. Therefore, if CIS genes are embedded in a region in which all genes are targeted at high frequency, they will not be considered significantly overtargeted by the regional Grubbs test for outliers.

In order to empirically evaluate and validate our method and the applied p-value corrections to stringently distinguish genotoxic CIS from local biases of integration, we

analyzed previously published vector integration datasets, in which known CIS originated by genetic selection are present. Moreover, we addressed whether the CIS identified by our method matched those identified with the more canonical statistical approach (49) We analyzed datasets from the γRV–based trials for X-linked Severe Combined Immunodeficiency (X-SCID) (17,18), Chronic Granulomatous Disease (CGD) (19), Wiskott Aldrich Syndrome (WAS) (12), and Adenosine Deaminase Severe Combined Immunodeficiency (ADA-SCID) (23, HSC-GT T-cell dataset), the LV-based trial for Adrenoleukodystrophy (ALD) (27) and the new LV-based trial for WAS object of this study. The analysis was done by pooling all vector integrations from all patients in each trial and, for the WAS clinical trial, the LV CIS were also analyzed in each single patient independently. The number of patients, integrations and targeted genes from the clinical trial integration datasets analyzed are reported in Table SOM13.

Table SOM13.

Clinical Trial	**N patients**	**N ISs**	**N targeted genes**
γRV X-SCID Paris	5	665	547
γRV X-SCID London	1	566	526
γRV CGD	2	677	493
γRV WAS	2	11,176	4,857
γRV ADA-SCID	4	957	756
LV ALD	2	2,610	1,805
LV WAS	3	23,139	7,119

Normality test of gene integration frequency distributions

The basic calculations for genome-wide Grubbs test are described in Biffi et al. manuscript (50). Briefly, the gene integration frequency is calculated by dividing the number of integrations targeting the same gene on genomic intervals defined by the boundaries of the genes (UCSC Hg19 freeze Jul 2012) targeted at least by one LV-integration. The gene integration frequency values are transformed by the minus logarithm base 2 to obtain a statistically evaluable normal distribution of the data. Indeed, the gene integration frequency values do not follow a normal distribution as they may vary from 0 to 1. The $-log_2$ transformed gene integration frequency shows that the data follow a normal distribution (D'Agostino & Pearson omnibus normality test) (Figure SOM13, Table SOM14).

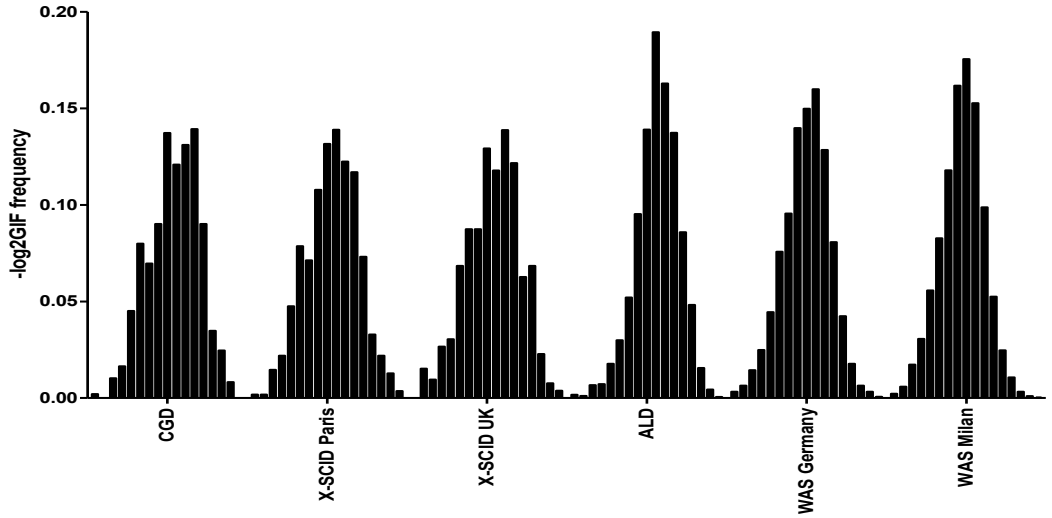


Figure SOM13.

Table SOM14.

D'Agostino & Pearson omnibus normality test	CGD	X-SCID Paris	X-SCID UK	ALD	WAS DE	WAS Milan
K2	4.8	5.0	4.0	2.9	3.3	2.8
P value	0.09	0.07	0.13	0.23	0.18	0.24
Passed normality test (alpha=0.05)?	Yes	Yes	Yes	Yes	Yes	Yes

For the MLD clinical trials in Milan the gene integration frequency was also analyzed separately for each patient. For Patient 1, 2 and 3 a total of 5455, 4084 and 4152 targeted genes were analyzed, respectively (Figure SOM14, Table SOM15).

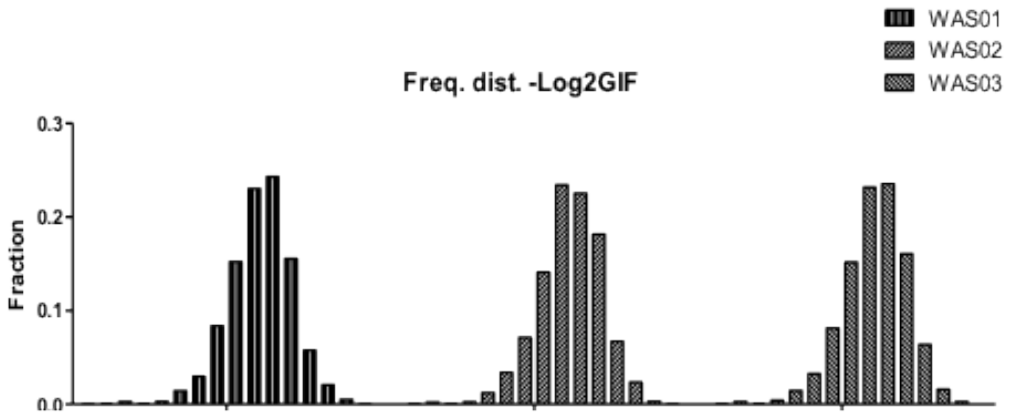


Figure SOM14

Table SOM15

D'Agostino & Pearson omnibus normality test	***Patient 1***	***Patient 2***	***Patient 3***
K2	5.0	4.5	4.7
P value	0.08	0.10	0.10
Passed normality test (alpha=0.05)?	Yes	Yes	Yes

Correction of CIS significance

To reduce the background of high integration frequency values originating by small genes, an additional 100 Kb distance was added to each gene interval. Thus, the Z-score, t-studentization and a raw p-value are calculated considering a minimal gene size of at least 100 Kb. Following the sequential p-value correction approach described above we first generated gene lists whose integration frequency provided a raw p value <0.05 (significant). The number of significant p values was subsequently used to multiply the raw p value of each gene and re-select for corrected p-values <0.05.

Before p-value correction, the number of significant CIS retrieved in each clinical trial correlated directly to the total number of targeted genes (R2=0.996). For datasets with <600 targeted genes (X-SCID and CGD) the significant CIS genes constituted on average the 1% of the entire dataset, while for datasets of >1800 targeted genes the CIS genes represented the 4% of the total. After p value correction, the number of CIS is reduced to the 0.2-0.5% of the total and the direct influence of the total number of targeted genes on the number of CIS identified vanishes. Thus, this correction allows performing a more stringent selection for the most over targeted genes.

A summary of the CIS identified in each clinical trial is shown on Table SOM16. The number of genes targeted by at least one vector integration varied depending on clinical trial analyzed (N Targeted Genes). The number of CIS identified before and after p-value correction is indicated in Figure SOM15 (N CIS raw p-value and N CIS corrected p-value respectively).

Table SOM16

Clinical Trial	***N Targeted Genes***	***N CIS raw p-value***	***N CIS corrected p-value***
γRV X-SCID Paris	547	8	3
γRV X-SCID London	526	3	0
γRV CGD	493	6	4
γRV WAS	4,857	212	7
γRV ADA-SCID	756	25	1
LV ALD	1,805	40	9
LV WAS	7119	312	7

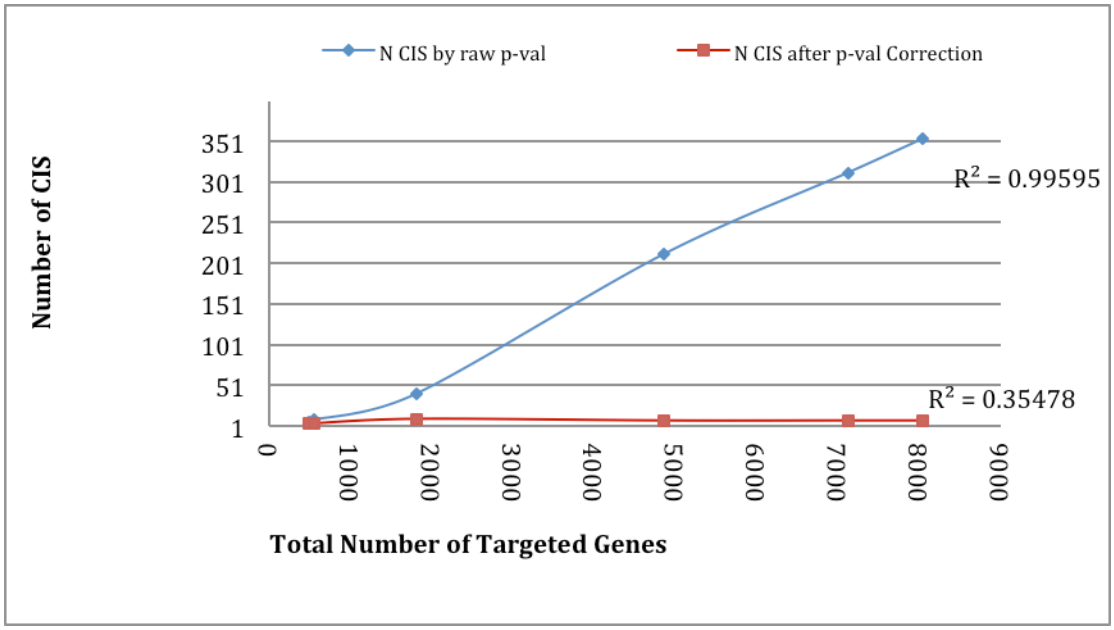


Figure SOM15.

CIS identification methods comparison

We then compared the CIS genes identified by the two methods. The comparison shows that the CIS identified by our method are a subset fully contained within the CIS dataset derived from Abel's regional analysis (Figure SOM16).

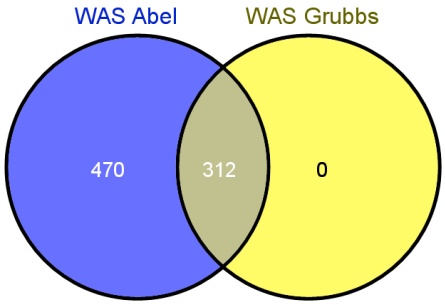


Figure SOM16

CIS identification in LV-based clinical trials by the genome-wide Grubbs test for outliers

The genome-wide Grubbs test for outliers applied to the previously reported ALD study and the present MLD integration datasets identified a number of significantly overtargeted LV CIS genes. Several of the CIS genes identified were shared between both trials (Figure SOM18). Given the larger size of the WAS dataset, it contained almost the entire ALD CIS dataset. After p-value correction, the number of CIS was reduced to 7-9 for each clinical trial. The most targeted CIS genes in both the ALD and WAS clinical trials were *KDM2A*, *TNRC6C*, *PACS1*, *C6orf10*, *MUM1*.

Table SOM17 summarize the CIS genes identified in the two LV-based clinical trials: Top 10 ranking significant genes are shown. The CIS genes with a p-value <0.05 after correction are in bold.

Table SOM17

Clinical Trial	CIS Genes (10 top ranking)
LV ALD	**PACS1 (18), TNRC6C (13), HLA-DQA1 (6), KDM2A (13), C6orf10 (10), MUM1 (6), SLITRK1 (5), HLA-DMB (5), SLITRK5 (5),** SLC22A11 (5)
LV WAS	**KDM2A (125), PACS1 (117), TNRC6C (94), NPLOC4 (78),** GRB2 (73), C6orf10 (70), ACOX1 (53), MUM1 (44), MIR612 (36), PPP1R2P1 (34)

We then assessed whether the CIS genes were significantly overtargeted when considering only the surrounding megabase-wide genomic region by the region-based Grubbs test. This analysis is required as the average gene integration frequency at the whole genome level does not take into account the biases of LV integration towards specific megabase-wide genomic regions. The gene integration frequency at specific chromosomal regions was visualized by plotting the p-values of gene integration frequencies along the gene chromosomal coordinates (Figure 2A of main manuscript). Local clusters of CIS within megabase-wide LV integration hotspots were evidenced by plotting CIS along the chromosomes. Given the skewed integration profile, 5 chromosomes contained ~50% of the LV CIS (chromosomes 11, 16, 17, 19) (Figure SOM17). All CIS from the WAS clinical trial were not considered significantly over targeted with respect to the neighboring genes by the region-based Grubbs test.

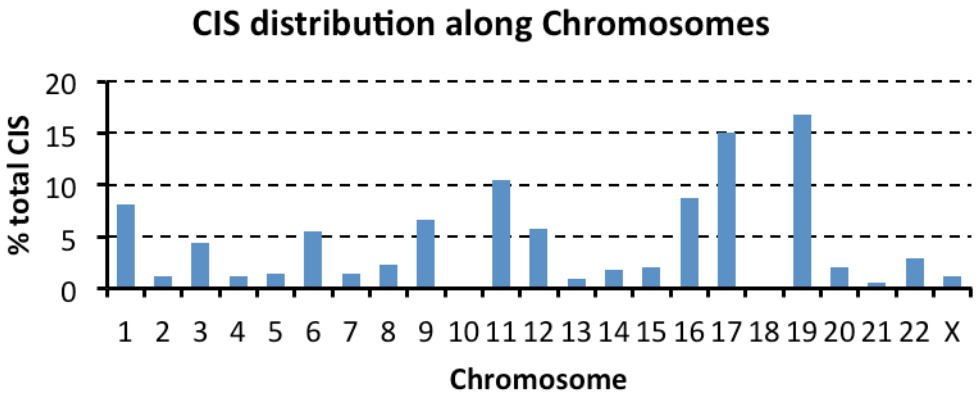


Figure SOM17.

CIS shared in HSC gene therapy trials before (**A**) and after (**B**) p-value correction are shown in Figure SOM18.

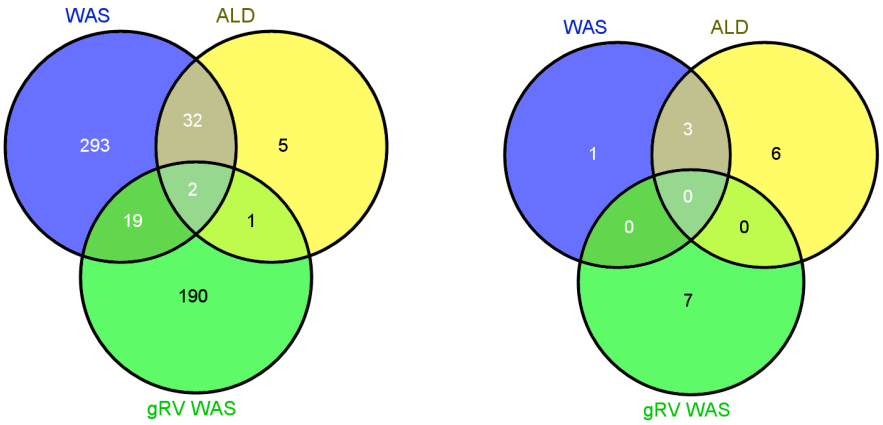


Figure SOM18

CIS identification in γRV-based clinical trials by the genome-wide Grubbs test for outliers

The genome-wide Grubbs test for outliers clearly identified, before and after p-value correction, CIS genes in γRV-based clinical trials such *LMO2*, *MECOM*, *PRDM16*, and *CCND2* together with other genes such as *ZNF217*, *TOMM20*, *KLF6*, *MIR17HG*, *MN1*, all being significantly over targeted (Supplementary Figure 30). This test however did not identify *SETBP1* or *RUNX1* as significant CIS, which were instead significant in previous analyses. Most CIS from these datasets were also previously analyzed by the regional Grubbs test for outliers (50). CIS significance evaluation considering the gene integration frequencies at the surrounding genomic intervals confirmed that well known culprits of oncogenesis/clonal-dominance such as *LMO2*, *CCND2*, *PRDM16*, *SETBP1*, *RUNX1* and *MECOM* were indeed significantly overtargeted in the X-SCID and CGD clinical trials. This finding confirms that our genome-wide Grubbs test for outliers is able to detect genotoxic CIS.

Table SOM18 shows the CIS genes identified in γRV-based clinical trials: Top 10 ranking significant genes are shown. In bold are indicated the CIS genes with a p-value <0.05 after correction.

Table SOM18.

Clinical Trial	CIS Genes (10 top ranking)
γRV X-SCID Paris	**ZNF217 (8), CCND2 (9), LMO2 (5),** TSRC1 (3), TOMM20 (3), FAM9C (3), PTGER4 (3), AFTIPHILIN (3)
γRV X-SCID London	FLJ20625 (3), TOMM20 (3), CD34 (3)
γRV CGD	**MECOM (91), PRDM16 (36), MN1 (5),** C6orf1 (3), NEUROD4 (3), TAOK3 (2)
γRV WAS	**LMO2 (33), MECOM (154), ZNF217 (25), TOMM20 (22), KLF6 (22), MIR17HG (20), CCND2 (21),** NRIP1 (30), FLI1 (18), MAP3K8 (16)
γRV ADA-SCID	**LMO2 (6),** ID2(4), CCDC77 (4), IRF2BP2 (4), CRTAM (5), C15ORF39 (4), ETS2 (4), MPG (5) HAAO (4), C10ORF95 (3)

As shown by the Venn diagrams in Figure SOM19, the most significant γRV CIS appear to be shared among different clinical trials.

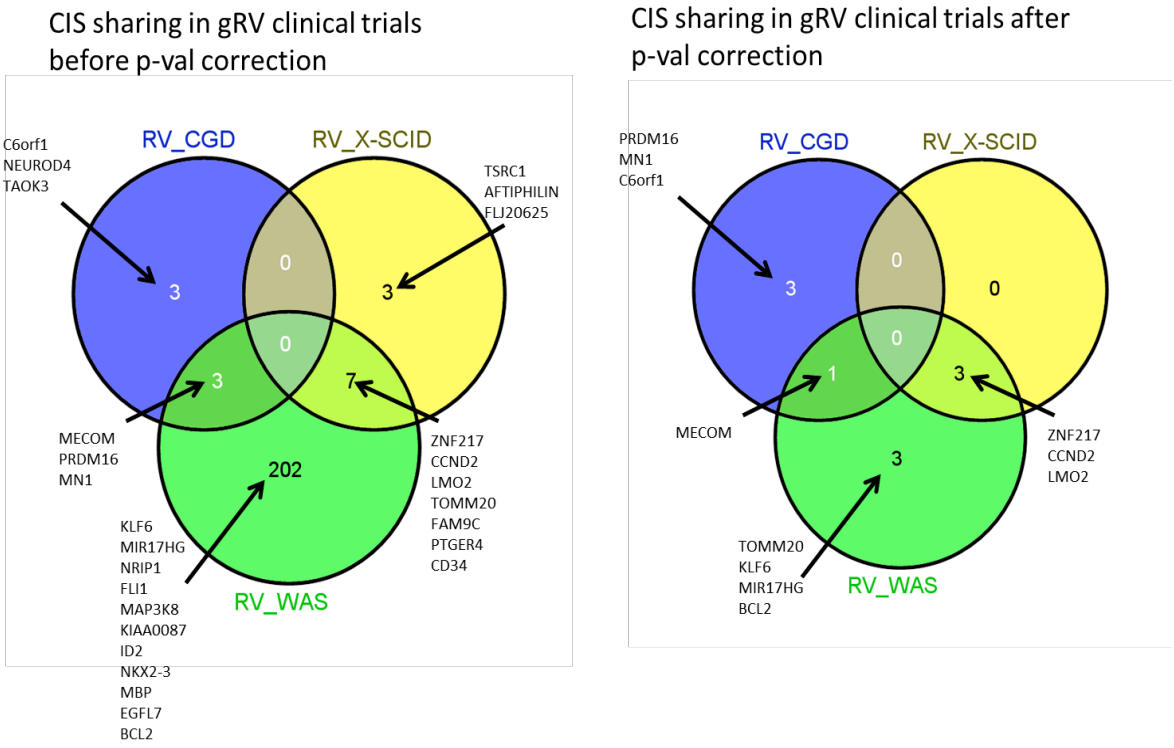


Figure SOM19.

Regional Grubbs test for outliers on selected CIS regions

Both the genome-wide methods used do not consider the local biases of integration. As a consequence CIS originating from vector related preferences would be highlighted. To verify whether the CIS from our clinical trial originate from local biases of integration we further analyzed a selected subset of CIS identified in our study by comparing their gene integration frequency to the neighboring genes (Figure SOM20).

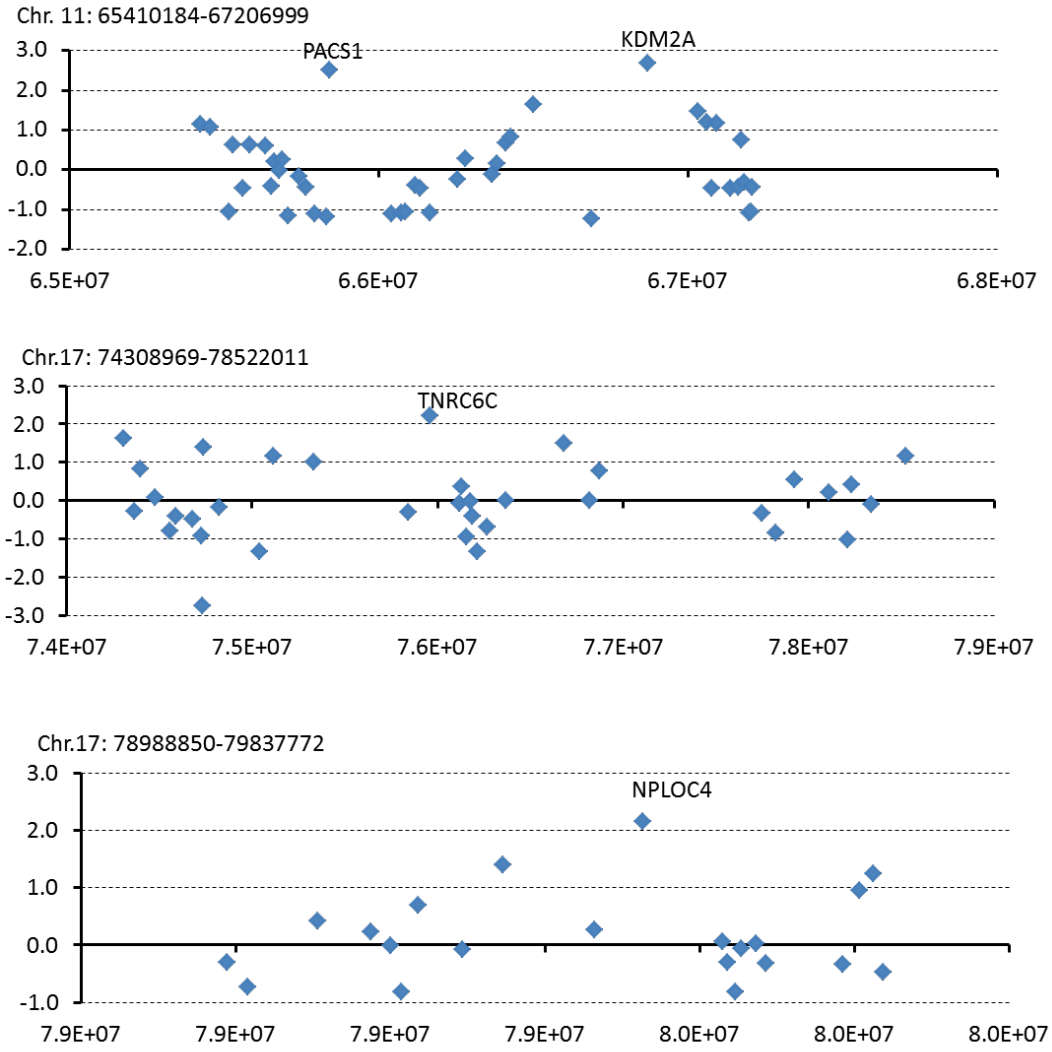


Figure SOM20.

Although some highlighted CIS carried a p-value <0.05 none of them reached the significance after Grubbs test. Complete data tables will be made available upon request.

Stem cell and Clonality analyses

Stem cells marking

Stem cells marking analysis was performed in order to retrieve a consistent group of ISs providing strong evidence of transduction and engraftment of multipotent progenitors. This information is relevant both for the efficacy of gene therapy and for general studies of hematopoietic cells biology after transplant. For this analysis we compared the following IS data sets from the indicated purified cell fractions taken as best representative of each hematopoietic subset/lineage: CD34+ cells purified from the BM as representative of the progenitors; purified PB CD3+ cells (T cells) and purified PB CD19+ cells (B cells), pooled to represent the Lymphoid lineage; purified PB CD14+ cells (granulocytes) and PB CD15+cells (monocytes), pooled to represent the Myeloid lineage.

The presence of shared IS between CD34+ cells, Myeloid and Lymphoid cells of the same patients was first assessed by comparing ISs after filtering only for collisions (shown for example for Patient 1 in Figure SOM21).

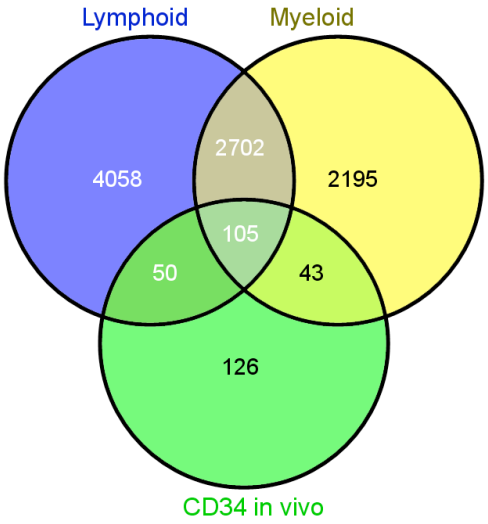


Figure SOM21.

To increase analytical stringency, we applied additional filtering levels. The first filter was based on sequence counts (≥3, as discussed above) resulting in the data shown in Figure SOM22 again for Patient1.

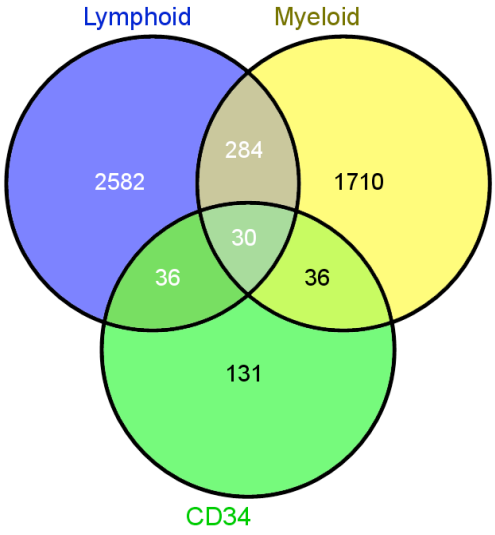


Figure SOM22.

We then applied a second filter taking into account the potential cross-contaminations among different samples from the same patient. Firstly, as discussed above in the Collision detection section, we estimated that background contamination between samples averages 0.02% in our experimental setup. Thus, any overlap of IS retrieved from different samples must well exceed this level for any further inference. Indeed, we observed much greater overlaps between the IS datasets retrieved from the different hematopoietic subtype/lineages of the same patient (as shown for Patient 1 in Figure SOM22). When interpreting these data, however, we must consider the purity of each cell fraction analyzed, which is dependent on the cell source, type and experimental methods used to purify that cell fraction. Based on empirical assessment, we considered that the extent of contamination by cells of any given different type/lineage in each purified cell fraction analyzed is likely to be less than 10%. Thus, we used the same filter previously designed to evaluate shared IS among different patients. Here, IS showing a sequence count 10 fold higher in a given lineage as compared to the others were assigned to the former lineage and discarded from the others. The total number of ISs remains the same as after the sequence count filter but the ISs are more stringently distributed among the lineages. Because the CD34+ cells fractions used for this analysis were only harvested from the BM and consistently purified to a purity approaching 90%, they are the cell population least likely to be contaminated by more than 10% of cells belonging to any given lineage. Moreover, they are the only fraction to comprise the progenitor cells. Thus, we based any further inference on stem cell marking only considering the IS shared between the CD34+ cell dataset and both the Lymphoid and Myeloid datasets. IS distributions after these filtering are shown in Figure SOM23. A list of the ISs putatively

representing stem/multipotent progenitor cells according to these criteria is available upon request.

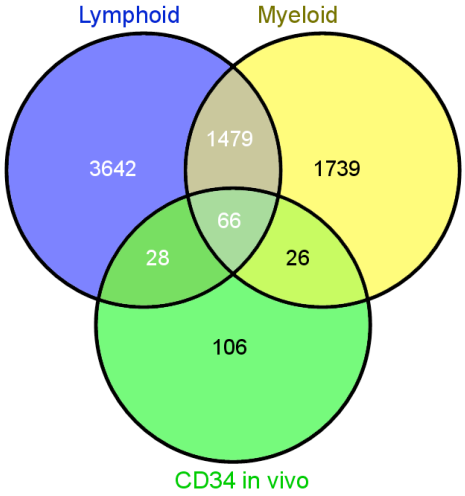


Figure SOM23.

Data derived from this filtering applied the other 2 patients are shown in Figure SOM24,25.

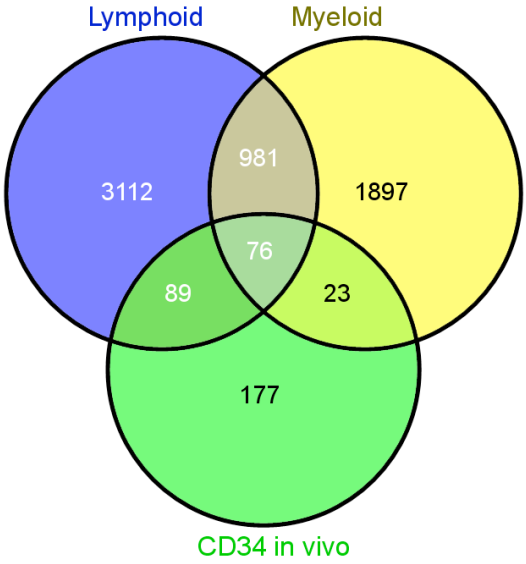


Figure SOM24.

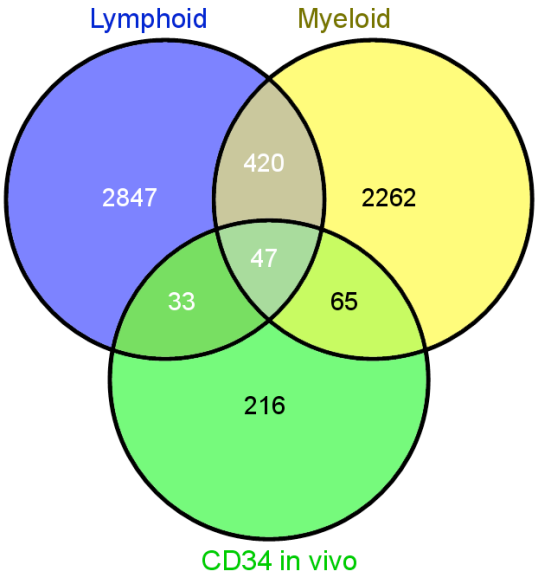


Figure SOM25.

Stem cell clones population size estimation

To compute a size estimation for stem cell population, we exploited patient 1 data, due to longer follow up, and we selected only short term lived Myeloid lineage from peripheral blood overtime as marker for stemness.

The estimate was performed through the mark-and-recapture approach using the Schnabel method, a modified version of the Petersen model (21, 46) that is summarized in the following formula:

$$N = \frac{(C_1 + 1)(C_2 + 1)}{S + 1} - 1$$

where N is the estimate of the population size, Ci is the number of captured elements at i-th time, and S is the number of shared elements between the two captures.

The Schnabel model requires:

- closed population, that means no births and no deaths/migrations of elements;
- equal probability of capturing among elements;
- independent sampling.

To best fit the Schnabel model requirements to stem cell clones' population, we selected time points from 9 months on, representing a stable condition, and we compared the two latest time points, 12 months versus 18 months.

Sampling ISs from these time points, we obtained the following counts (Table SOM19):

Sampling	Model Var	Captured ISs
12	C1	558
18	C2	1612
shared	S	141

Table SOM19

Thus the Schnabel estimator became:

$$N = \frac{(C_1 + 1)(C_2 + 1)}{S + 1} - 1 = 6348$$

achieving an estimate of around 6300 elements as population size of stem cell clones.

To overcome and balance possible model biases due to requirements for equal probability of capturing among elements, we also tested these findings exploiting a mark-recapture model that can assess and take into account probability heterogeneity of capture. Using Poisson regression models (log-linear) (46) we compared 3 available samplings (time points 9, 12 and 18 months) and we obtained an estimate of 12125 HSC, resulting in higher but within the same order of magnitude than the PS model.

Clonality and Diversity analysis

The analysis of clonal diversity addresses the issue of IS diversification in the different lineage compartments. The Shannon information content index provides a measurement of entropy and diversity of each of our IS datasets, as shown in a previous work by (21) using the formula

$$H' = -\sum_{i=1}^{R} p_i \log p_i$$

where $\boldsymbol{p_i}$ is the proportion of sequence counts belonging to the *i*th type of IS in a given cell lineage. Results of this analysis in each patient are shown in Figure 2 of the main manuscript.

SOM 5: Histone modifications distribution around IS

In order to study the chromatin features surrounding each IS we profiled the density distribution of Histone modifications (47) in the genomic areas involved by vector integration events. The analysis shown in Figure 3 of the main manuscript is performed in the same fashion as previously published (54). We also analyzed specific loci where relative frequency of each histone modification is shown in bins of 2kb in size on frequency distribution graphs. ISs frequency in the same loci is shown for the same bins size on reverse histograms in grey (Figure S16-18).

SOM 6: Gene Ontology analysis

Gene Ontology analysis (GO) on ISs data from the 3 MLD patients was performed through GREAT (http://great.stanford.edu) a statistical tool originally developed for datasets from chromatin immunoprecipitation studies.

As compared to other gene-based platforms standardly used for GO analysis, such as Ingenuity Pathway Analysis (http://www.ingenuity.com) or DavidEASE (http://david.abcc.ncifcrf.gov/) (which have been originally designed for microarray studies), GREAT performs both a hypergeometric test of ontological enrichment and a binomial test of enrichment for genomic regions surrounding each gene. GREAT accepts genomic positions as input and weights the results according to the presence of insertional hotspots in given genomic regions and to the functions of the targeted genes. The readout is a list of GO classes resulted significant both for the binomial and for the hypergeometric test. We used GREAT as enrichment analysis tool on the in vitro and in vivo integration datasets for each MLD patient independently. Moreover, for comparison purposes we also analyzed the LV integrations retrieved after gene therapy from 2

patients of the previously reported ALD clinical trial (27). The analysis was performed considering the single nearest gene within 1 Mb from each IS mapped on human genome assembly hg19. As threshold for significance we set a False Discovery Rate (FDR) <0.05 for both the hypergeometric and binomial tests mentioned above.

To measure the degree of similarity between the overrepresented GO classes in the MLD and ALD in vivo ISs datasets, we calculated the percentage of gene sharing (***V***) as follows: given two classes c_1 and c_2, ***O*** the number of overlapping genes of the two classes, and ***S*** the total number of genes belonging to the class with the lowest number of genes between c_1 and c_2 (Figure S19-20):

$$V = \frac{O}{S} * 100$$

More detailed data will be made available upon request.

Supplementary Figures Legends

Figure S1

Flow chart of transduced CD34+ cell manufacturing.

Figure S2

Flow chart for LV WAS supernatant production. MCB: master 293T cell bank; CF10: 10-tray cell factory.

Figure S3

WASp expression. (A)Transduced and untransduced CD34+ cells from Pt1 were cultured for 15 days in the presence of cytokines and WASp expression was analysed by Western Blot. (B) Western Blot was performed on ex-vivo isolated PB MNC, EBV transformed B cell line and untransformed T cell line generated in vitro 1 year post-GT from Pt2. PBMC and EBV transformed B cell line from a healthy donor are shown as control.

Figure S4

VCN per genome evaluated by qPCR at different time points after GT (up to 2.5 years) in the erythroid lineages purified from the BM of the 3 WAS patients.

Figure S5

WASp expression in different PB population after GT. Flow cytometric characterization of WASp expression in combination with specified antibodies was performed on PB T, B, NK cells and monocytes and on platelets was performed before GT and at different

time points after GT. Representative staining of pre-GT (Pt2) and post-GT (2.5 years Pt1 and 1.5 year for Pt2 and Pt3) and a healthy donor are shown.

Figure S6

WASp expression in BM and PB-derived platelets. WASp expression was analysed both on PB (A) and BM (B) derived platelets and Pt1 and Pt2 (1 year after GT) and a healthy donor (HD).

Figure S7

T, B and NK cell count during follow up. Absolute counts of T (CD3+, CD4+ and CD8+), B and NK cells are reported before and at different time points after GT. Period of patients' treatment with busulfan (Bu), Fludarabine (Flu), Rituximab (anti-CD20mAb) and steroid (methylprednisolone, MPD) during the conditioning and after GT are indicated on the graphs.

Figure S8

Proliferation after TCR activation in Pt1 and Pt3 T-cell lines. T-cell lines from healthy donors, Pt1 (A) and Pt3 (B) were labeled with CFSE and tested for the proliferative capacity in response to different concentration of anti-CD3 mAb. The percentage of CFSE dilution is shown.

Figure S9

Polyclonal TCR Vβ repertoire. 24 different TCR Vβ specificities were analysed by flow

cytometry before (white) and at latest follow up (2.5 years Pt1 and 1.5 year for Pt2 and Pt3) after gene therapy (yellow). The average values for healthy donors are shown in orange. The percentage of each TCR Vbeta on CD3+ cells is shown.

Figure S10

WASp expression in CD4+/CD25bright cells after gene therapy in Pt1 and Pt3. WASp expression was measured by flow cytometry analysis in Pt1 and Pt3 at 6 and 12 months after GT, respectively.

Figure S11

Restoration of immunological synapses after gene therapy. Immunological synapses formation was evaluated on Pt1 cells and a healthy donor by the expression of WASp, actin and perforin at the interface between the NK effector cells and the target K562 cells. Stained cells were analysed by confocal microscopy.

Figure S12

Representative samples Spreadex gel electrophoresis of Linear Mediated Amplification PCR (LAM PCR) reactions for 3 WAS patients. LAM PCR products at latest FU. Internal control bands are indicated as IC.

Figure S13

Representative samples Spreadex gel electrophoresis of Linear Mediated Amplification PCR (LAM PCR) reactions for 3 WAS patients. LAM PCR products for different lineages overtime. Internal control bands are indicated as IC

Figure S14

Genomic distribution of LV vs RV vector. A) Analyses of genomic distribution on *in vitro* and *in vivo* pooled data from 3 LV GT patients vs *in vivo* pooled data from 2 RV GT patients. Graphs on the left show on the X-axis the IS distance with respect to TSS of genes and on the Y axis the relative IS frequency. Graphs on the right shows on the X axis the number of genes present in a +/-10kb window surrounding IS and on the Y axis the relative IS frequency; B) Percentage of IS landing inside genes in each RV (Retro) and LV (Lenti) GT patient; C) number of IS retrieved in PRDM16, CCND2 and HMGA2 loci from each patient. The graphical representation is identical to lower boxes of Figure 3F from main manuscript.

Figure S15

A) Genomic distribution of ISs from in vitro and in vivo pooled dataset of our 3 LV GT treated patients. Chromosomes' size is reported on top of the graph. Refseq Genes and ISs frequency distributions are shown in bins of 1Mbp (grey, coloured columns respectively). B) Incidence of ISs from LV and γ-retro GT in each chromosome. Analysis as performed for Figure 4E of the main manuscript (*$p<0.05$ Fisher exact test).

Figure S16

Frequency of IS and histone marks in the MECOM region. ISs frequencies from RV and LV GT are shown in bins of 2kb on reverse histograms in grey. Frequencies of histone modifications in the same bins are shown in colored histogram below.

Figure S17

Frequency of IS and histone marks in the STAT5A-B/STAT3 region. ISs frequencies from RV and LV GT are shown in bins of 2kb on reverse histograms in grey. Frequencies of histone modifications in the same bins are shown in colored histogram below.

Figure S18

Frequency of IS and histone marks in the KDM2A region. ISs frequencies from RV and LV GT are shown in bins of 2kb on reverse histograms in grey. Frequencies of histone modifications in the same bins are shown in colored histogram below.

Figure S19

Chromatin modifications surrounding ISs. Probability density distributions of histone modifications mapped on $CD34^+$ cells in a +/-45kb window surrounding IS are shown as heatmaps. Color intensity show under (blue) or over (yellow) representation of each given histone modification as compared to random *in silico* generated reference. Each row represents an *in vitro* or *in vivo* ISs dataset from the 3 LV and 2 γ-retro GT treated patients. Analysis performed as in Figure 4B.

Figure S20

Heatmap showing similarities between “Molecular Functions” GO classes (from 0% [white] to 100% [red] of overlapping genes) that were significantly hit in WAS GT (rows) vs ALD GT (column).

Figure S21

Heatmap showing similarities between “Biological Processes ” GO classes (from 0% [white] to 100% [red] of overlapping genes) that were significantly hit in WAS GT (rows) vs ALD GT (column).

Supplementary Figures:

Figure S1

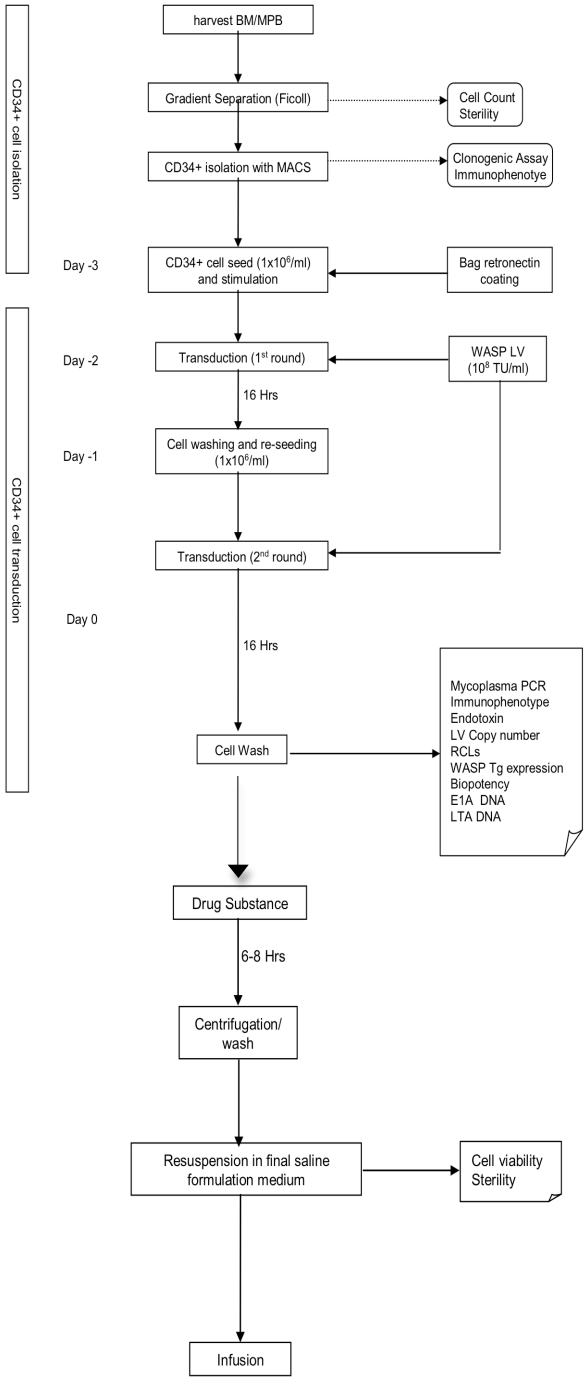


Figure S2

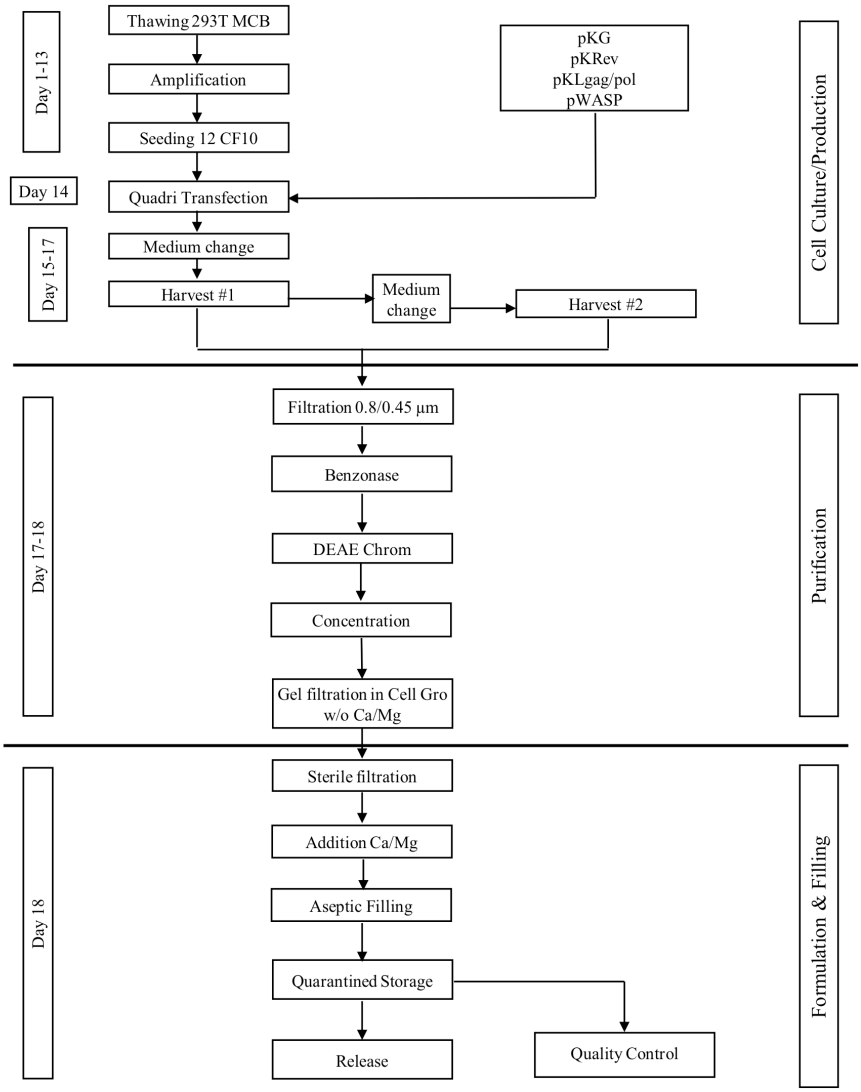

Day 1-13
Thawing 293T MCB
Amplification
Seeding 12 CF10
pKG
pKRev
pKLgag/pol
pWASP
Day 14
Quadri Transfection
Day 15-17
Medium change
Harvest #1
Medium change
Harvest #2
Cell Culture/Production
Day 17-18
Filtration 0.8/0.45 µm
Benzonase
DEAE Chrom
Concentration
Gel filtration in Cell Gro w/o Ca/Mg
Purification
Day 18
Sterile filtration
Addition Ca/Mg
Aseptic Filling
Quarantined Storage
Release
Quality Control
Formulation & Filling

Figure S3

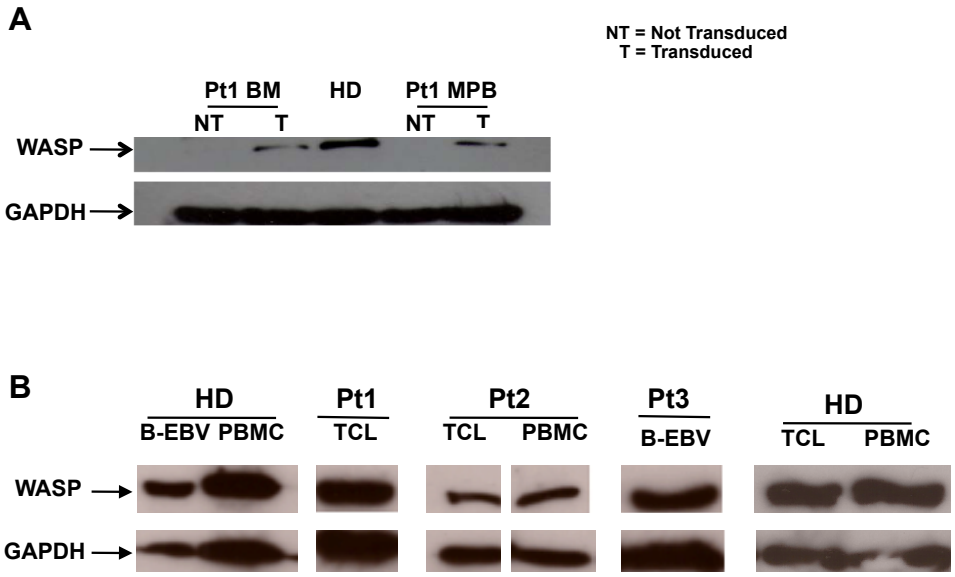

A
NT = Not Transduced
T = Transduced
Pt1 BM
HD
Pt1 MPB
NT
T
NT
T
WASP
GAPDH
B
HD
B-EBV PBMC
Pt1
TCL
Pt2
TCL
PBMC
Pt3
B-EBV
HD
TCL
PBMC
WASP
GAPDH

Figure S4

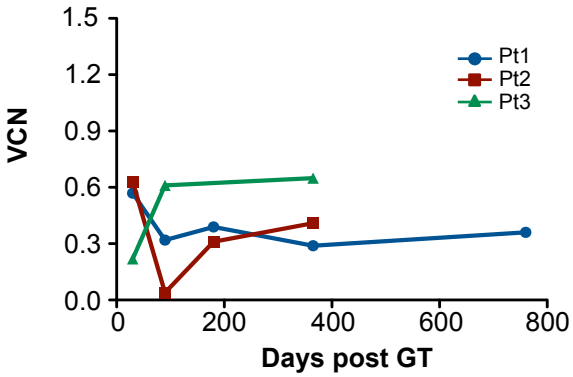


Figure S5

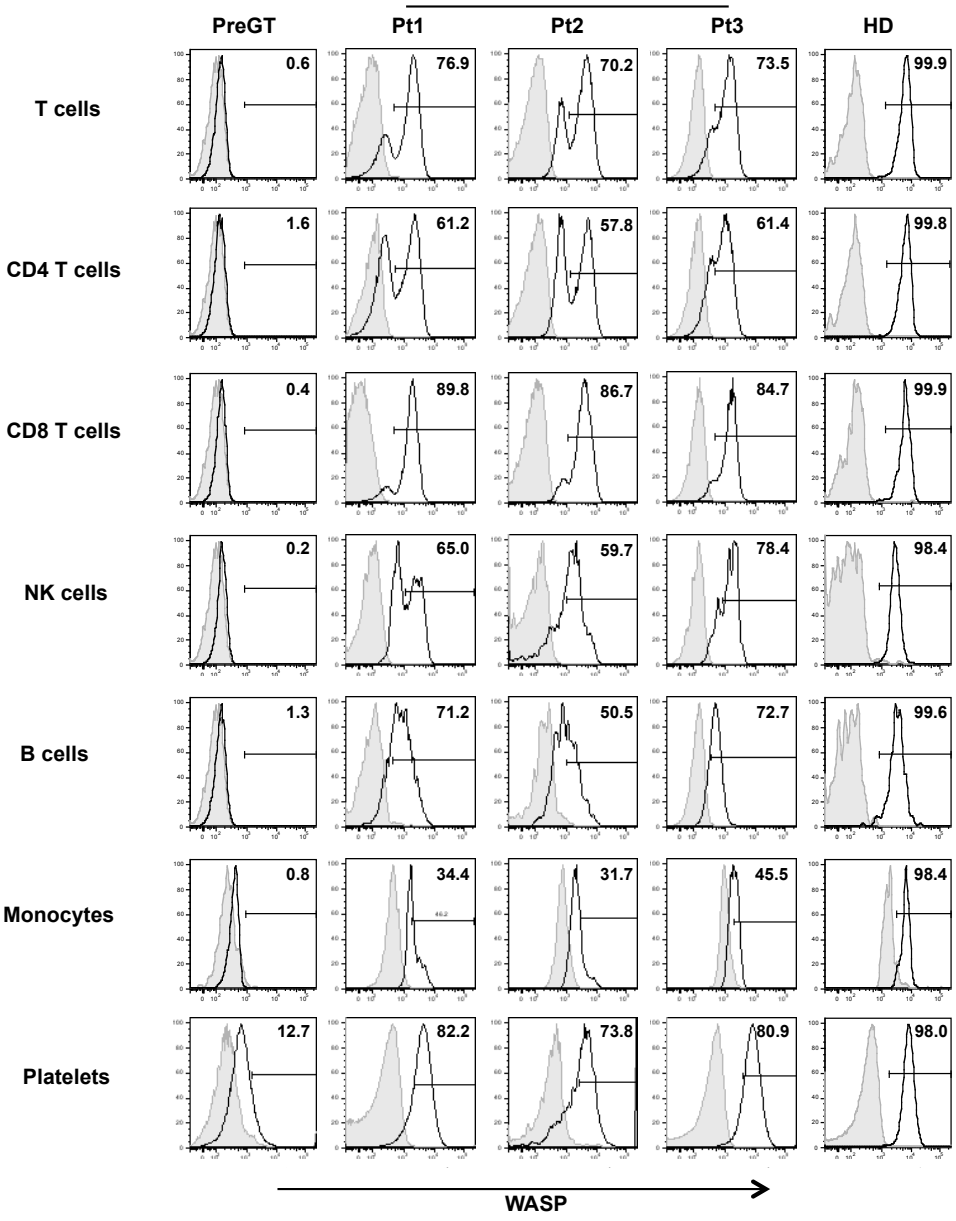

PreGT
Pt1
Pt2
Pt3
HD
T cells
0.6
76.9
70.2
73.5
99.9
CD4 T cells
1.6
61.2
57.8
61.4
99.8
CD8 T cells
0.4
89.8
86.7
84.7
99.9
NK cells
0.2
65.0
59.7
78.4
98.4
B cells
1.3
71.2
50.5
72.7
99.6
Monocytes
0.8
34.4
31.7
45.5
98.4
Platelets
12.7
82.2
73.8
80.9
98.0
WASP

Figure S6

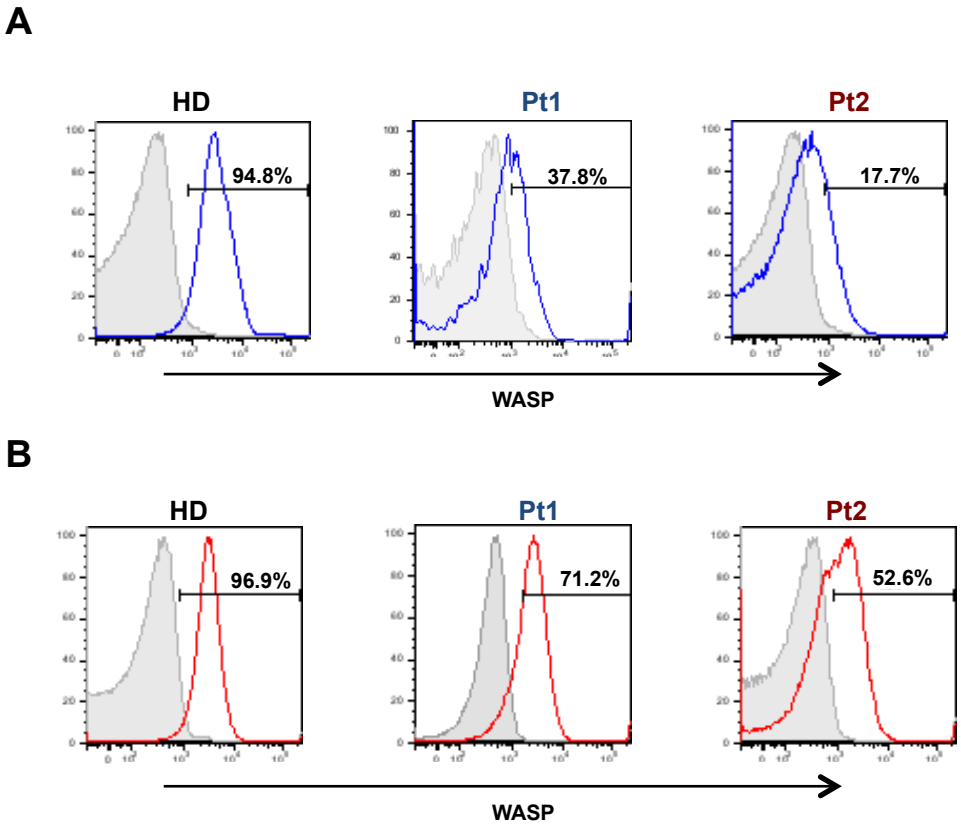


Figure S7

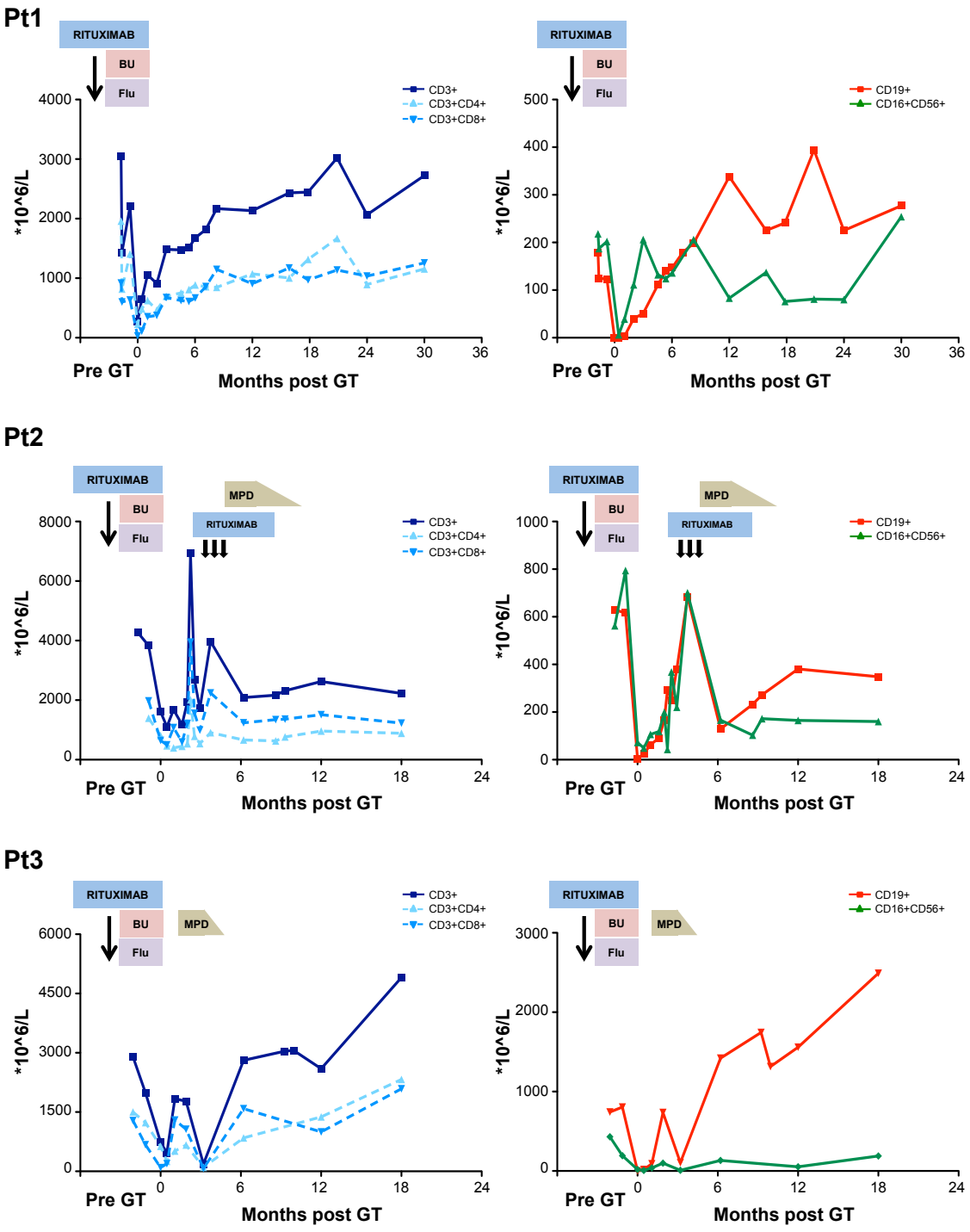

Pt1
RITUXIMAB
BU
Flu
CD3+
CD3+CD4+
CD3+CD8+
*10^6/L
Pre GT
Months post GT
CD19+
CD16+CD56+
Pt2
MPD
Pt3

Figure S8

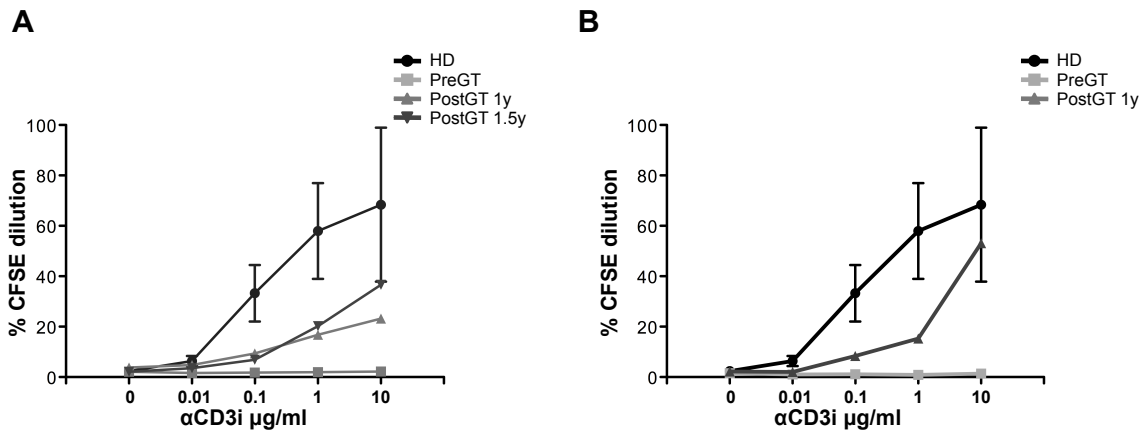

A
HD
PreGT
PostGT 1y
PostGT 1.5y
% CFSE dilution
100
80
60
40
20
0
0
0.01
0.1
1
10
αCD3i µg/ml
B
HD
PreGT
PostGT 1y
% CFSE dilution
100
80
60
40
20
0
0
0.01
0.1
1
10
αCD3i µg/ml

Figure S9

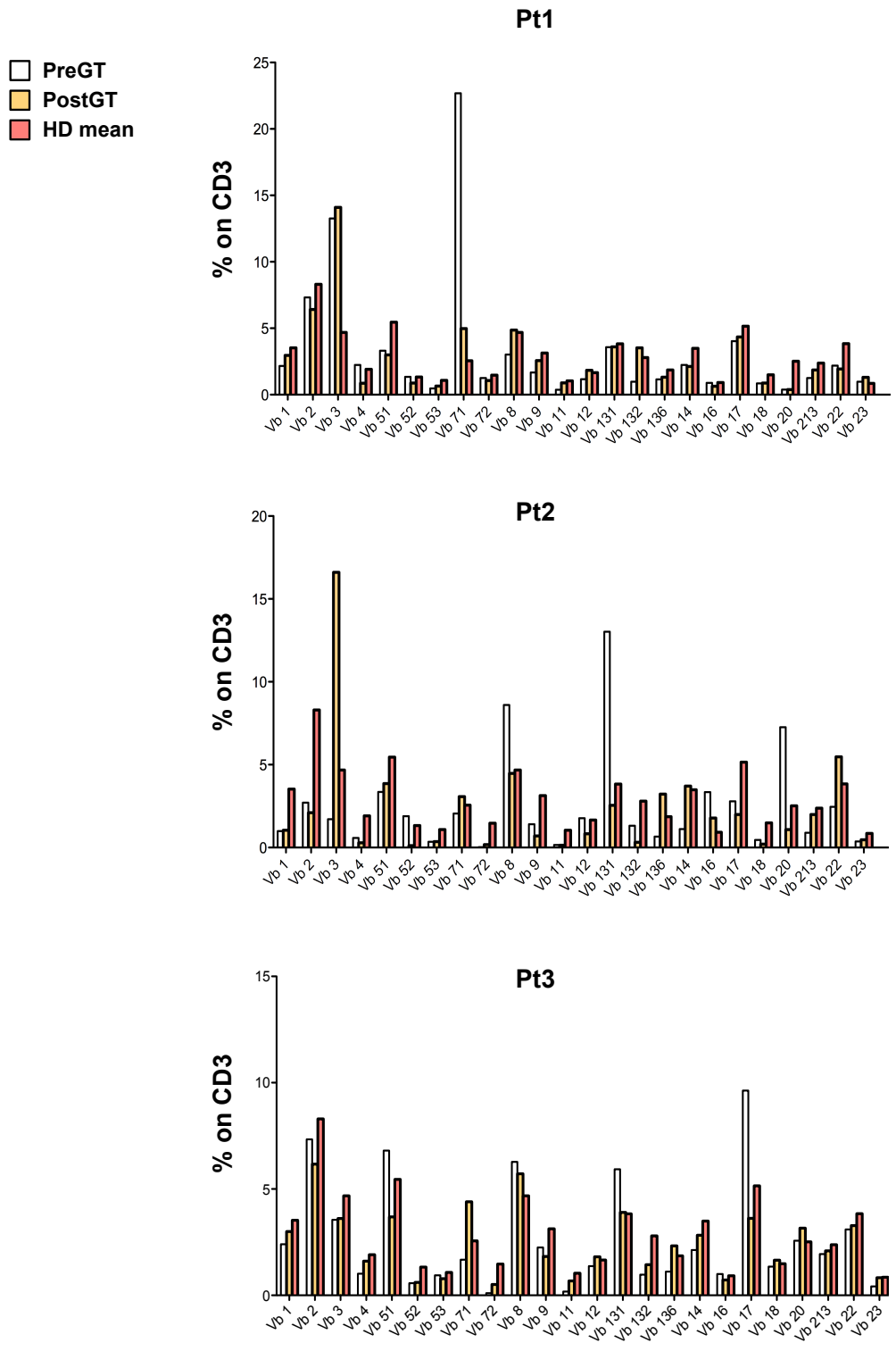

PreGT
PostGT
HD mean
Pt1
Pt2
Pt3
% on CD3
0
5
10
15
20
25
Vb 1
Vb 2
Vb 3
Vb 4
Vb 51
Vb 52
Vb 53
Vb 71
Vb 72
Vb 8
Vb 9
Vb 11
Vb 12
Vb 131
Vb 132
Vb 136
Vb 14
Vb 16
Vb 17
Vb 18
Vb 20
Vb 213
Vb 22
Vb 23

Figure S10

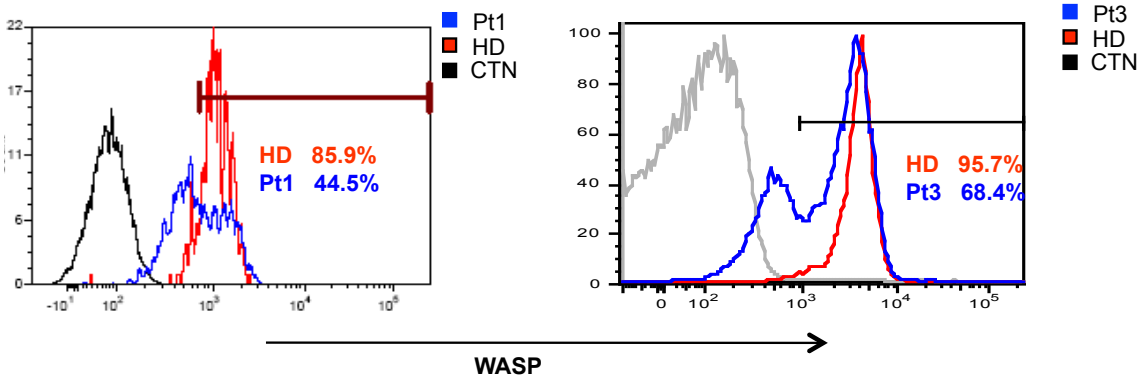
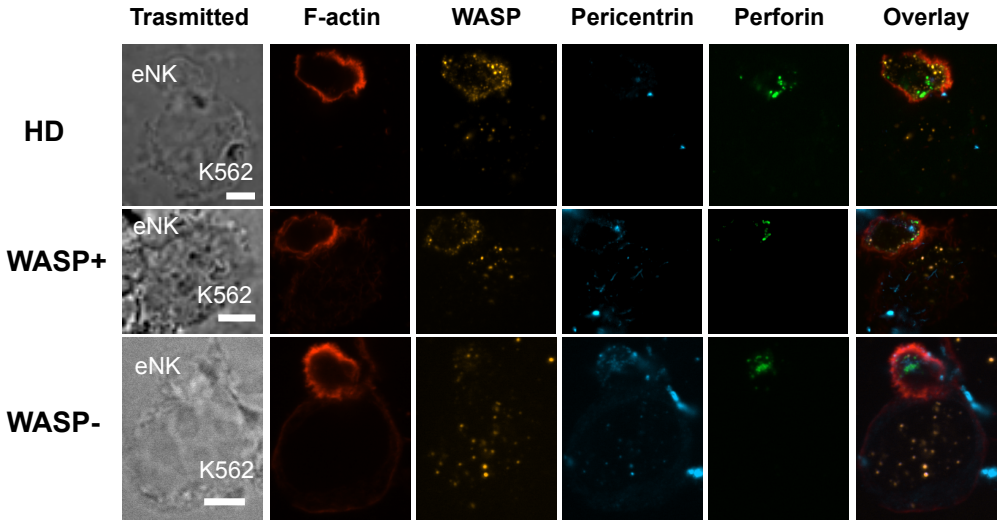


Figure S11



Scale bar= 5μm

Figure S12

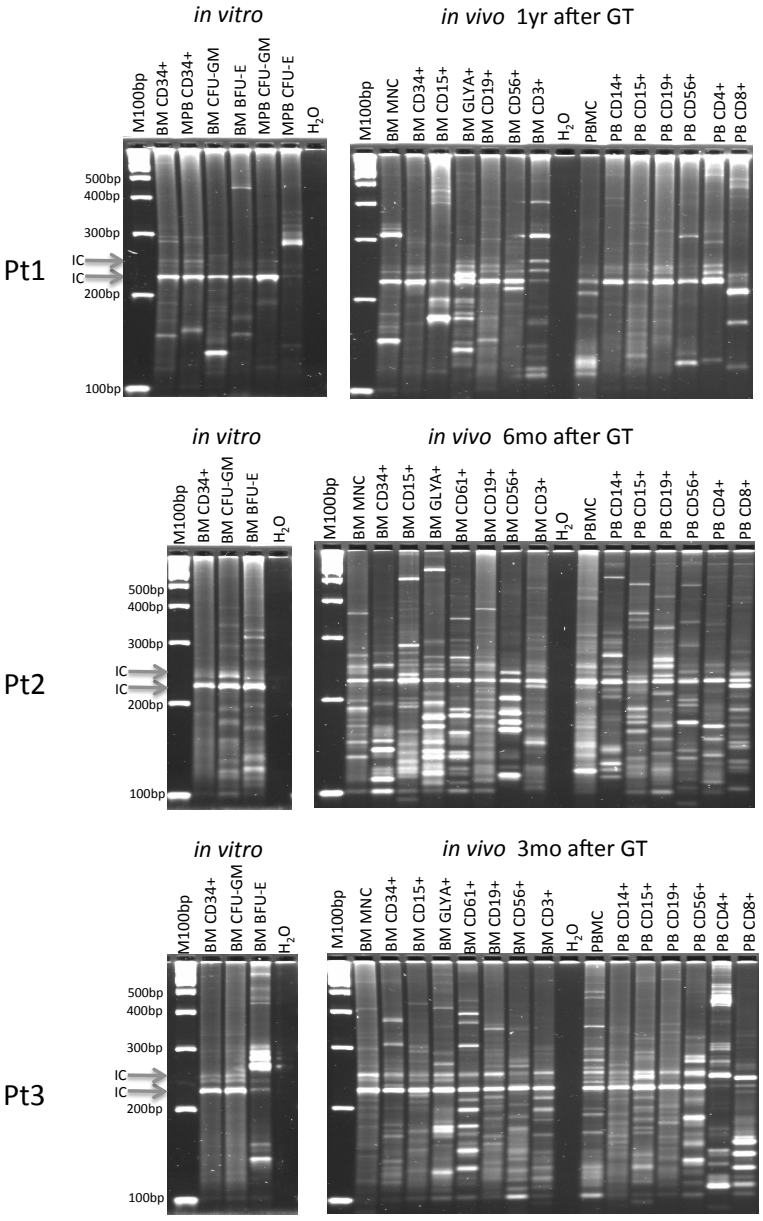

in vitro
in vivo 1yr after GT
M100bp
BM CD34+
MPB CD34+
BM CFU-GM
BM BFU-E
MPB CFU-GM
MPB CFU-E
H2O
BM MNC
BM CD15+
BM GLYA+
BM CD19+
BM CD56+
BM CD3+
PBMC
PB CD14+
PB CD15+
PB CD19+
PB CD56+
PB CD4+
PB CD8+
500bp
400bp
300bp
IC
IC
200bp
100bp
Pt1
in vitro
in vivo 6mo after GT
BM CD61+
Pt2
in vitro
in vivo 3mo after GT
Pt3

Figure S13

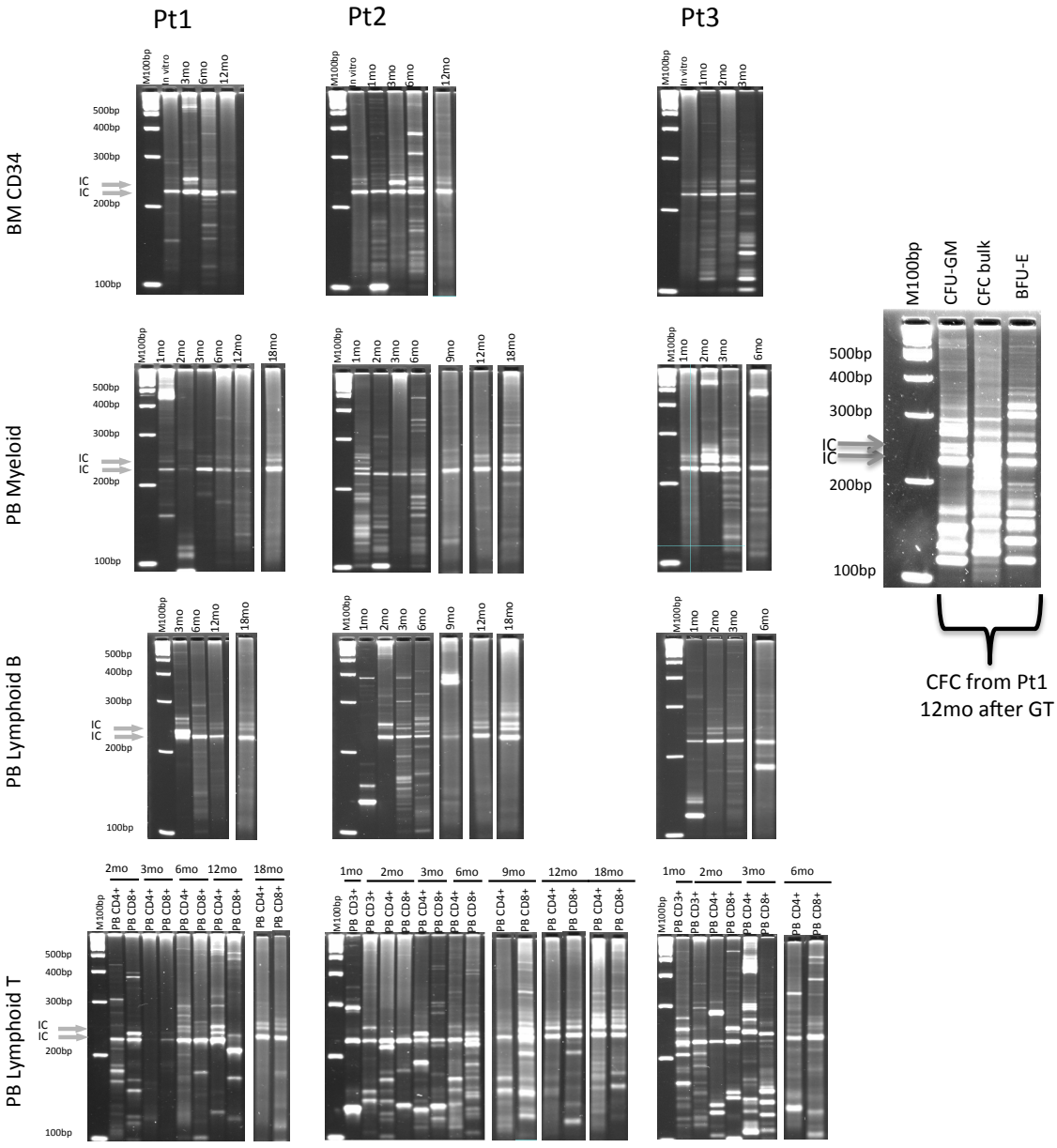

Pt1
Pt2
Pt3
BM CD34
PB Myeloid
PB Lymphoid B
PB Lymphoid T
M100bp
In vitro
1mo
2mo
3mo
6mo
9mo
12mo
18mo
PB CD4+
PB CD8+
PB CD3+
500bp
400bp
300bp
IC
200bp
100bp
CFU-GM
CFC bulk
BFU-E
CFC from Pt1
12mo after GT

Figure S14

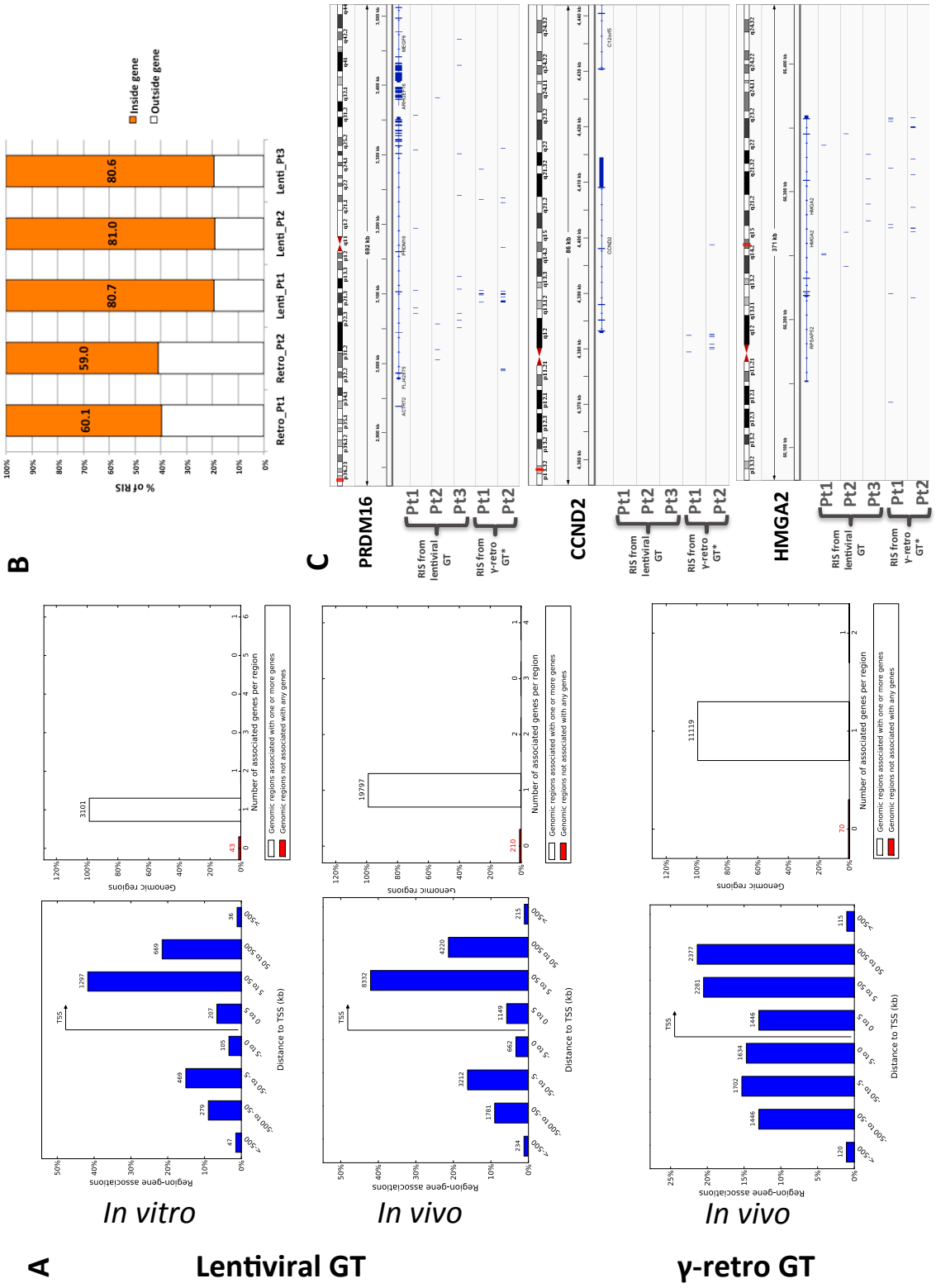

A
Lentiviral GT
γ-retro GT
In vitro
In vivo
In vivo
Region-gene associations
Distance to TSS (kb)
Genomic regions
Number of associated genes per region
Genomic regions associated with one or more genes
Genomic regions not associated with any genes
B
% of RIS
Inside gene
Outside gene
Retro_Pt1
Retro_Pt2
Lenti_Pt1
Lenti_Pt2
Lenti_Pt3
60.1
59.0
80.7
81.0
80.6
C
PRDM16
CCND2
HMGA2
RIS from lentiviral GT
RIS from γ-retro GT*
Pt1
Pt2
Pt3

Figure S15

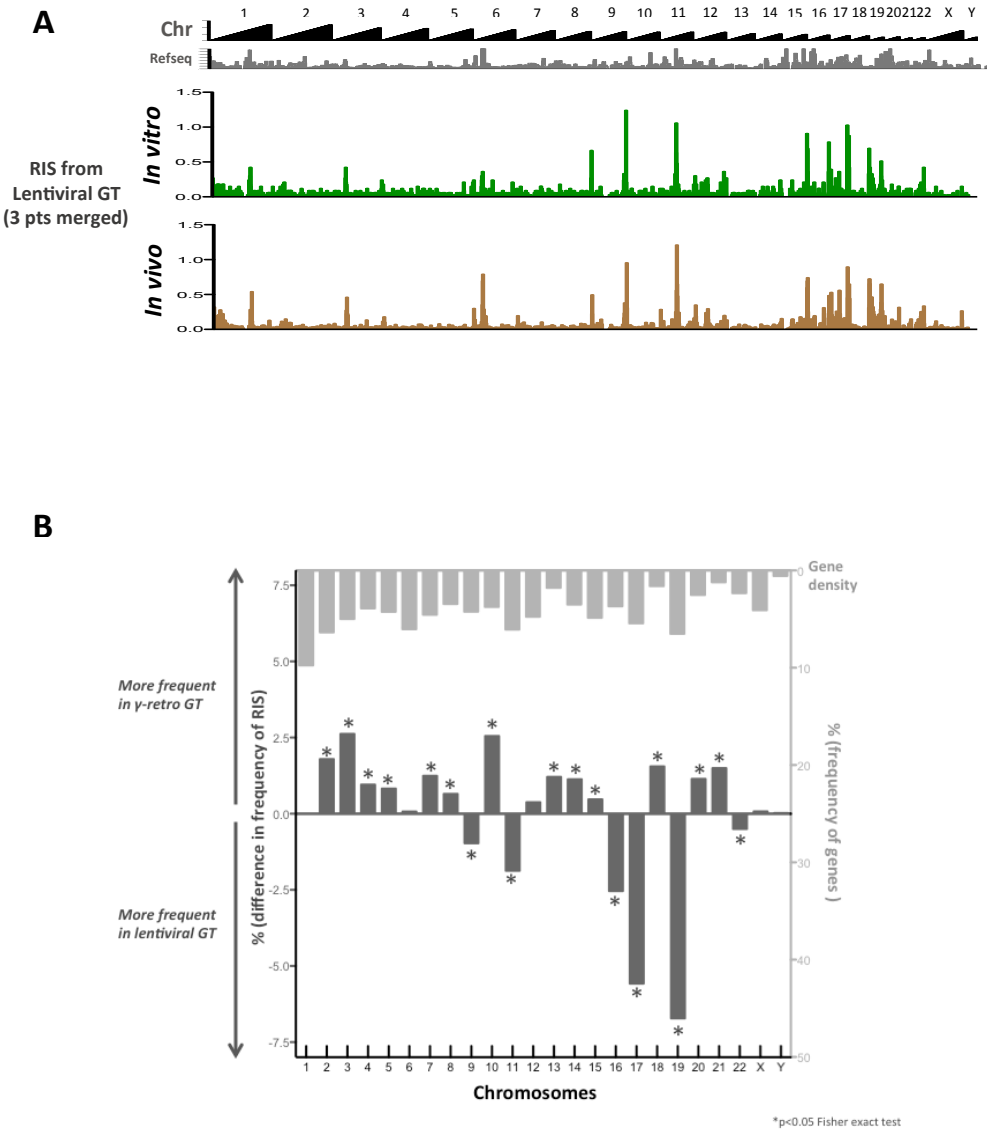

A
Chr
Refseq
RIS from
Lentiviral GT
(3 pts merged)
In vitro
In vivo
B
More frequent
in γ-retro GT
More frequent
in lentiviral GT
% (difference in frequency of RIS)
Gene
density
% (frequency of genes)
Chromosomes
*p<0.05 Fisher exact test

Figure S16

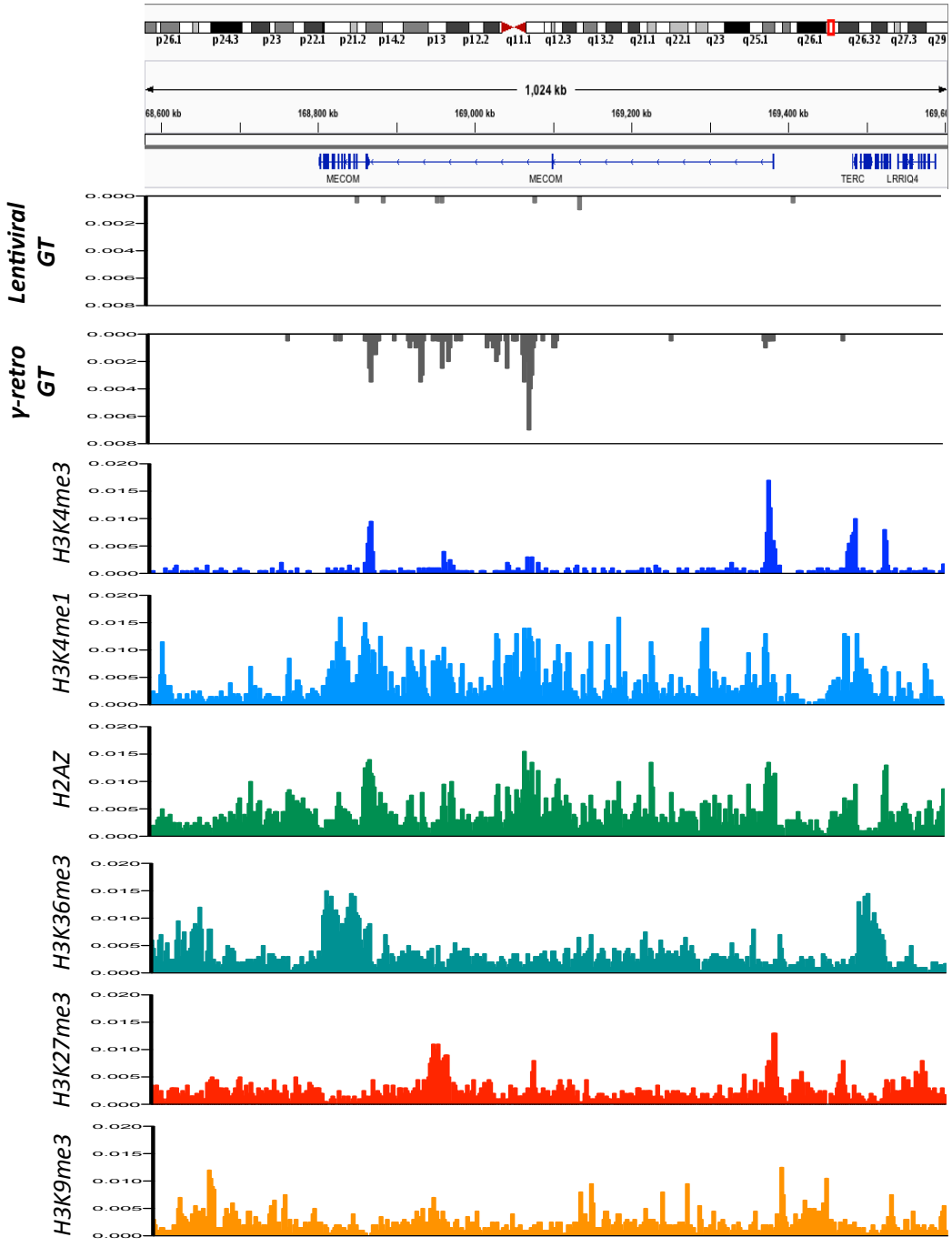

p26.1 p24.3 p23 p22.1 p21.2 p14.2 p13 p12.2 q11.1 q12.3 q13.2 q21.1 q22.1 q23 q25.1 q26.1 q26.32 q27.3 q29
1,024 kb
168,600 kb
168,800 kb
169,000 kb
169,200 kb
169,400 kb
MECOM
MECOM
TERC
LRRIQ4
Lentiviral GT
γ-retro GT
H3K4me3
H3K4me1
H2AZ
H3K36me3
H3K27me3
H3K9me3
0.000
0.002
0.004
0.006
0.008
0.010
0.015
0.020
0.005

Figure S17

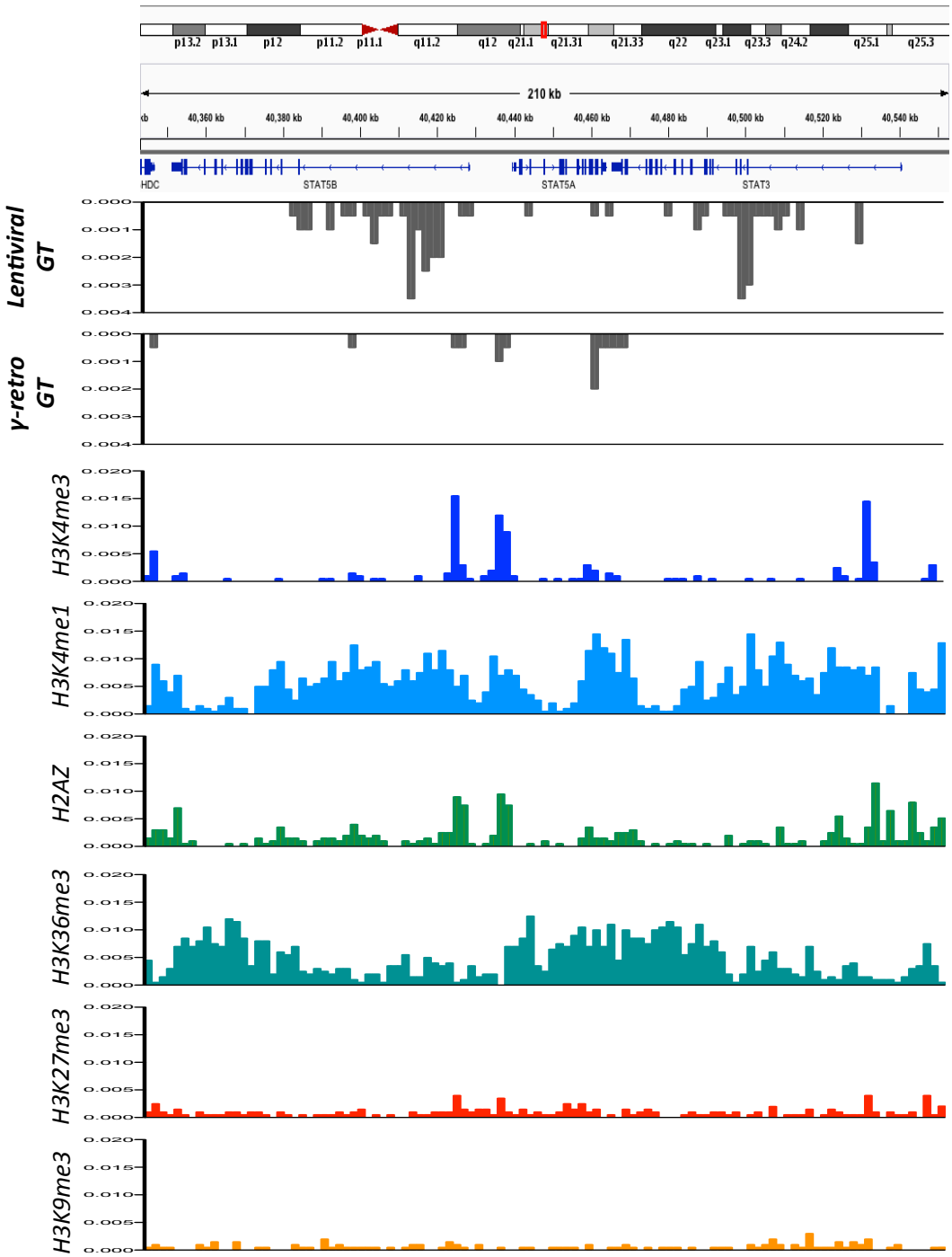

p13.2 p13.1 p12 p11.2 p11.1 q11.2 q12 q21.1 q21.31 q21.33 q22 q23.1 q23.3 q24.2 q25.1 q25.3
210 kb
40,360 kb 40,380 kb 40,400 kb 40,420 kb 40,440 kb 40,460 kb 40,480 kb 40,500 kb 40,520 kb 40,540 kb
HDC STAT5B STAT5A STAT3
Lentiviral GT
γ-retro GT
H3K4me3
H3K4me1
H2AZ
H3K36me3
H3K27me3
H3K9me3

Figure S18

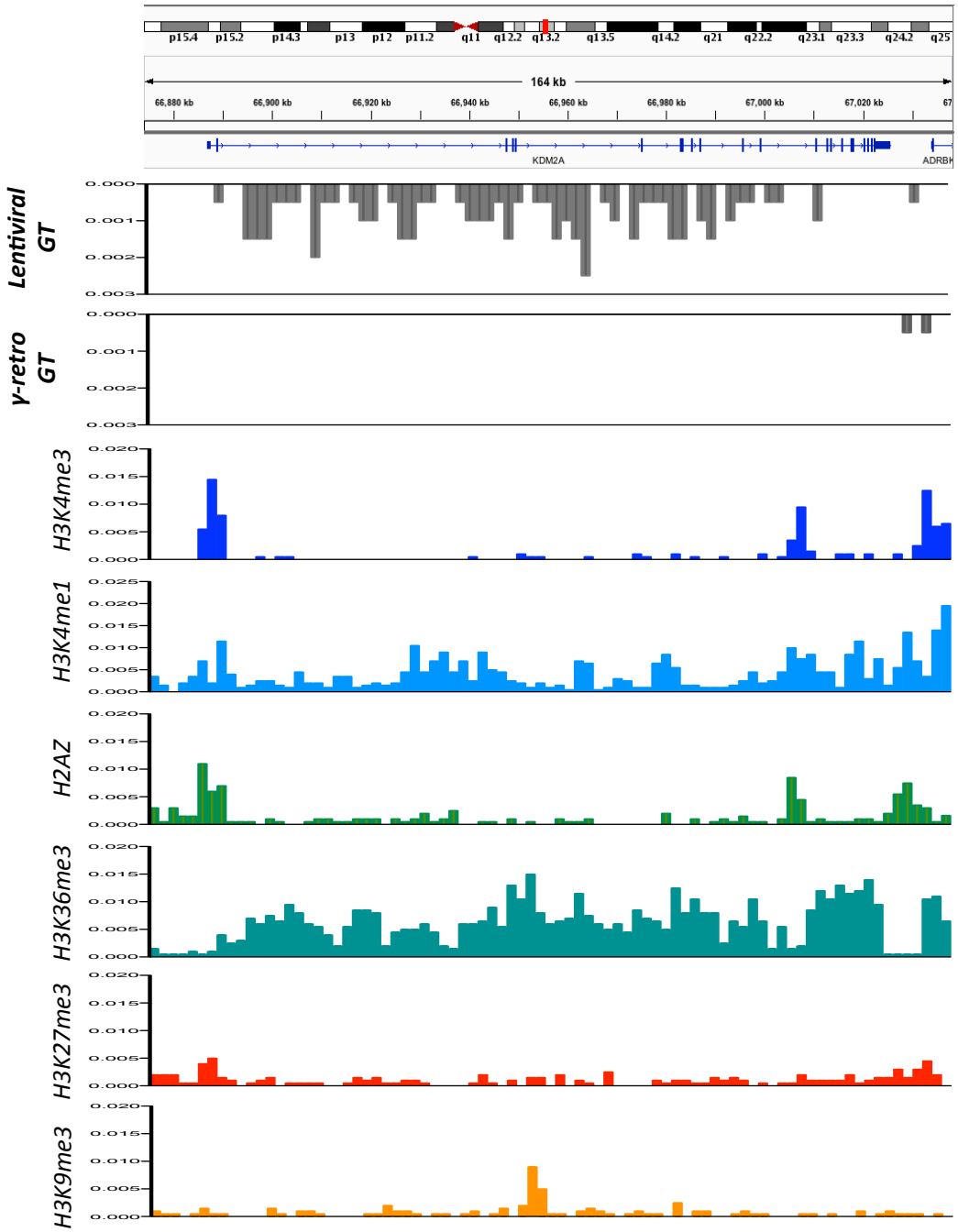

p15.4 p15.2 p14.3 p13 p12 p11.2 q11 q12.2 q13.2 q13.5 q14.2 q21 q22.2 q23.1 q23.3 q24.2 q25
164 kb
66,880 kb 66,900 kb 66,920 kb 66,940 kb 66,960 kb 66,980 kb 67,000 kb 67,020 kb
KDM2A
ADRBK
Lentiviral GT
γ-retro GT
H3K4me3
H3K4me1
H2AZ
H3K36me3
H3K27me3
H3K9me3

Figure S19

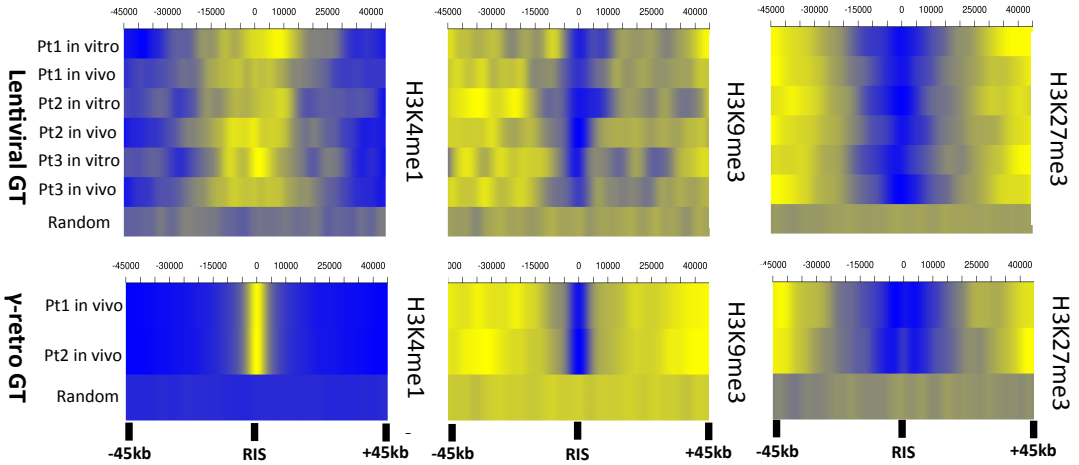

Lentiviral GT
Pt1 in vitro
Pt1 in vivo
Pt2 in vitro
Pt2 in vivo
Pt3 in vitro
Pt3 in vivo
Random
γ-retro GT
Pt1 in vivo
Pt2 in vivo
Random
H3K4me1
H3K9me3
H3K27me3
-45000 -30000 -15000 0 10000 25000 40000
-45kb
RIS
+45kb

Figure S20

WAS \ ALD		histone methyltransferase activity	histone-lysine N-methyltransferase activity	protein-lysine N-methyltransferase activity	protein methyltransferase activity	nuclear hormone receptor binding	hormone receptor binding	steroid hormone receptor binding	protein N-terminus binding	DNA helicase activity	helicase activity	ATP-dependent helicase activity	ligand-dependent nuclear receptor binding	phosphatase binding	mitogen-activated protein kinase kinase kinase binding
		50	40	41	67	109	126	66	83	48	154	115	19	108	16
histone methyltransferase activity	50	50	40	39	50	2	2	2	2	0	0	0	1	0	0
histone-lysine N-methyltransferase activity	40	40	40	38	40	1	1	1	2	0	0	0	1	0	0
protein-lysine N-methyltransferase activity	41	39	38	41	41	1	1	1	2	0	0	0	1	0	0
protein methyltransferase activity	67	50	40	41	67	2	2	2	2	0	0	0	1	0	0
N-methyltransferase activity	66	48	40	41	53	2	2	2	2	0	0	0	1	0	0
androgen receptor binding	39	2	1	1	2	39	39	39	4	0	2	1	4	1	0
nuclear hormone receptor binding	109	2	1	1	2	109	109	66	5	0	4	3	9	3	0
hormone receptor binding	126	2	1	1	2	109	126	66	5	0	4	3	9	3	0
steroid hormone receptor binding	66	2	1	1	2	66	66	66	4	0	4	3	7	3	0
ligand-dependent nuclear receptor transcription coactivator activity	42	1	0	0	1	23	23	9	1	0	0	0	8	0	0
protein N-terminus binding	83	2	2	2	2	5	5	4	83	4	5	3	3	3	2
DNA helicase activity	48	0	0	0	0	0	0	0	4	48	48	38	0	0	0
ATP-dependent DNA helicase activity	36	0	0	0	0	0	0	0	3	36	36	36	0	0	0
signaling adaptor activity	67	0	0	0	0	1	2	1	2	0	0	0	0	3	4
SH3/SH2 adaptor activity	50	0	0	0	0	0	1	0	1	0	0	0	0	1	0
ARF guanyl-nucleotide exchange factor activity	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0
oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors	49	0	0	0	0	0	0	0	0	0	0	0	0	0	0

0% 1-25% 26-50% 51-75% 76-99% 100%

Figure S21

WAS \ ALD		chromatin modification	histone modification	covalent chromatin modification	histone lysine methylation	histone methylation	protein methylation	histone H3-K27 methylation	regulation of transcription elongation from RNA polymerase II promoter	nuclear transport	nucleocytoplasmic transport	T cell costimulation	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	DNA repair	interspecies interaction between organisms	nuclear migration	establishment of nucleus localization	nucleus localization	muscarinic acetylcholine receptor signaling pathway
		420	264	269	48	61	79	2	12	327	324	79	25	397	398	7	11	19	12
histone modification	264	264	264	264	48	61	61	2	7	12	11	0	0	32	27	0	0	0	0
covalent chromatin modification	269	269	264	269	48	61	61	2	7	13	12	0	0	32	27	0	0	0	0
histone lysine methylation	48	48	48	48	48	48	48	2	2	0	0	0	0	3	1	0	0	0	0
macromolecule methylation	139	67	64	67	45	56	70	2	3	3	3	1	0	9	6	0	0	0	0
regulation of gene expression, epigenetic	129	49	32	37	12	15	15	2	0	6	6	1	0	17	6	0	0	0	0
gene silencing	85	29	17	22	5	5	5	0	0	7	7	0	0	7	6	0	0	0	0
nuclear transport	327	16	12	13	0	0	0	0	2	327	324	5	0	14	57	0	0	2	0
nucleocytoplasmic transport	324	15	11	12	0	0	0	0	2	324	324	5	0	13	56	0	0	2	0
nuclear export	112	3	2	2	0	0	0	0	0	112	112	0	0	3	16	0	0	1	0
mRNA export from nucleus	64	0	0	0	0	0	0	0	0	63	63	0	0	2	11	0	0	1	0
antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	25	0	0	0	0	0	0	0	0	0	0	15	25	0	1	0	0	0	0
T cell costimulation	79	0	0	0	0	0	0	0	0	5	5	79	15	0	13	1	1	1	0
DNA repair	397	47	32	32	3	4	5	0	0	14	13	0	0	397	19	0	0	0	0
G2/M transition DNA damage checkpoint	32	8	5	5	1	1	1	0	0	1	1	0	0	19	1	0	0	0	0
G2/M transition checkpoint	41	8	5	5	1	1	1	0	0	1	1	0	0	24	2	0	0	0	0
interspecies interaction between organisms	398	38	27	27	1	4	4	0	1	56	56	13	1	19	398	0	0	0	0
response to ionizing radiation	108	23	16	16	3	3	3	0	0	7	7	1	0	46	10	0	0	0	0
Golgi vesicle transport	179	1	1	1	0	0	1	0	0	2	2	1	0	1	5	0	0	0	0
regulation of defense response to virus	57	1	1	1	0	0	0	0	0	7	7	8	0	0	20	0	0	0	0
regulation of response to biotic stimulus	74	2	2	2	0	0	0	0	0	7	7	8	0	0	25	0	0	0	0

0% 1-25% 26-50% 51-75% 76-99% 100%

Table S1

Process results for WAS GMP LV batches produced used in this study

PARAMETER	GMP BATCHES		
	08073	*08078*	*09010*
Initial titer (TU/ml)	$1.2x10^7$	$1.7x10^7$	$1.0x10^7$
Initial Physical particles (ng p24/ml)	$3.0x10^2$	$3.5x10^2$	$3.6x10^2$
Initial Infectivity (TU/ng p24)	$4.0x10^4$	$5.0x10^4$	$3.0x10^4$
Initial Volume (Litres)	25	25	25
Final titer (TU/ml)	$7.6x10^8$	$5.0x10^8$	$2.1\ x10^8$
Final Physical particles (ng p24/ml)	$1.2x10^4$	$1.3x10^4$	$8.6x10^3$
Final Infectivity (TU/ng p24)	$6.1x10^4$	$3.8x10^4$	$2.4\ x10^4$
Final Volume (ml)	180	180	180
Overall Yield (TU)	45%	21%	14%
Overall Yield (p24)	29%	27%	17%
Total TU	$1.4x10^{11}$	$9.0x10^{10}$	$3.8x10^{10}$

Table S2

Characterization of the GMP WAS LV batches produced

TEST	SPECIFICATION	BATCH NO.		
		08073	*08078*	*09010*
Physicochemical and Identity				
Osmolality (mOsm/Kg)	260-350	303	304	305
pH EP 2.2.3	7.0-8.0	7.6	7.7	7.7
WASP transgene sequence	Corresponding	Corresponding	Corresponding	Corresponding
Vector integrity	Corresponding to reference	Corresponding	Corresponding	Corresponding
Lentiviral proteins	Corresponding to reference	Corresponding	Corresponding	Corresponding
Potency and Bioactivity				
InfectiousTiter (TU/ml)	$\geq 2.0 \times 10^8$	7.6×10^8	5.0×10^8	2.1×10^8
Physical titer (HIV Gag p24 Antigen) (ng/ml)	FIO	1.2×10^4	1.3×10^4	8.6×10^3
Infectivity (Transducing unit/ng p24)	$\geq 2.0 \times 10^4$	6.1×10^4	3.8×10^4	2.4×10^4
Transgene expression	WAS protein presence	Present	Present	Present
Microbial purity and safety				
Sterility EP 2.6.1	Negative	Negative	Negative	Negative
Mycoplasma EP 2.6.7 (cultural assay)	Negative	Negative	Negative	Negative
Endotoxin EP 2.6.14 (quantitative assay) (EU/ $2x10^8$ TU)	≤ 25	14	21	9
In vitro Adventitious viruses	Negative	Negative	Negative	Negative
In vivo Adventitious viruses	Negative	Negative	Negative	Negative
RCL	Negative	Negative	Negative	Negative
Process and product impurities				
Host cell proteins (ng/ $2x10^8$ TU)	FIO	21	33	79
Plasmid residual DNA (VSV-G) (copies/ $2x10^8$ TU)	$\leq 4x10^8$	0.4×10^8	0.7×10^8	2.3×10^8
Large T antigen protein contamination (ng/ml)	$\leq$ LLOQ	$\leq$ LLOQ	$\leq$ LLOQ	$\leq$ LLOQ

Large T antigen Residual DNA (copies / $2x10^8$ TU)	$\leq 2x10^5$	1.1 x 104	1.9 x 104	9.2 x 104
Benzonase contamination (ng/ml)	≤ 0.2	< 0.1	< 0.1	< 0.1
E1A residual DNA (copies/ $2x10^8$ TU)	$\leq 2x10^5$	2.0x 104	4 .0x 104	9.1 x 104
Total residual DNA (µg/ $2x10^8$ TU)	FIO	0.5	0.9	4
BSA contamination (µg/ $2x10^8$ TU)	FIO	0.3	0.5	1.5
Vector cross contamination	$\leq 10^5$pp/10^{10} pp	$\leq 10^5$pp/10^{10} pp	$\leq 10^5$pp/10^{10} pp	$\leq 10^5$pp/10^{10} pp

FIO: for information only; RCL: replication competent lentivirus; LOQ: limit of quantification.

Table S3

Additional information on patients' treatment

	Pt1	**Pt2**	**Pt3**
$CD34^+$ cells collected (purified) ($x10^6$/kg)	4.45 (BM) + 6.54 (MPB*)	14.5	13.0
% of $CD34^+$ cells in infused cells	66.1% (BM) 96.2% (MPB)	82.7%	89.6%
Compliance release criteria	Yes	Yes	Yes

* After thawing.

Table S4

Tests and specifications for Drug Substance and Drug Product quality control

TEST	SPECIFICATION	AVAILABLE AT INFUSION
Mycoplasma PCR (*EofT*) *	Negative	yes
LV Copy number *(LC)*	FIO	no
WASP Transgene product expression *(LC)*	WAS protein presence	no
RCL (EofT)	Negative	no
Large T antigen DNA (*EofT*) and (*LC*)	*EofT:* FIO	yes
	LC : ≤ LLOQ	no
E1A DNA (*EofT*) and (*LC*)	*EofT:* FIO	yes
	LC : ≤ LLOQ	no
Endotoxin* *(EofT)*	≤ 2.5 EU /ml	yes
Clonogenic test *(EofT)*	FIO	no
Transduction efficiency *(EofT)*	FIO	no
Immunophenotype *(EofT)* (CD34, CD45, CD19, CD3, CD15)	FIO	yes
Sterility – Microbiological control of cellular products (Drug product)	Negative	no
Cell viability (Drug Substance)	≥ 80%	yes

EofT: end of transduction; LC: 14 days liquid culture; FIO: for information only; RCL: replication competent lentivirus; LLOQ: lower limit of quantification; IF: Immunophenotyping.

Table S5

Chemotherapy and neutropenia

	Pt1	Pt2	Pt3
Total dose of Busulfan (mg/kg)	7.6	9.4	10.1
Cumulative Busulfan AUC (ng*h/ml)	46072	33314	42980
Days of neutropenia (ANC < 500/µl)	From d+13 to d+24 (12 days)	From d+8 to d+24 (17 days)	From d+11 to d+29 (19 days)
Neutropenia Nadir (ANC and day)	N=0, day+17	N=200, day+12	N=0, day+14

ANC, absolute neutrophil counts; AUC, area under the curve.

Table S6

Samples sources analyzed at each timepoint

Timepoint (months)		Pt1	Pt2	Pt3
in vitro	0 (transduction)	BM/MPB CD34	BM CD34	BM CD34
in vivo	1	BM; PB	BM; PB	BM; PB
	2	PB	PB	BM; PB
	3	BM; PB	BM; PB	BM; PB
	6	BM; PB	BM; PB	
	12	BM; PB	-	-

(BM= Bone Marrow, PB= Peripheral Blood, MPB= Mobilized peripheral blood)

Table S7

Samples analyzed for vector integration studies :Genomic DNA extracted from whole Bone Marrow (BM) and Peripheral Blood (PB) or FACS-sorted cell sub-types from these tissues, harvested from patients at different time points.

	Tissue	Cell Type	Marker
in vitro	BM	In vitro cultured	CD34+
		Colonies CFC*	-
in vivo	PB	PBMC§	-
		Lymphoid T	CD3+, CD4+, CD8+
		Myeloid	CD14+, CD15+
		Lymphoid B	CD19+
		Lymphoid NK	CD56+
	BM	BM MNC§	-
		Lymphoid T	CD3+
		Myeloid	CD15+
		Lymphoid B	CD19+
		CD34+ cells	CD34+,
		Lymphoid NK	CD56+
		Erythroid	CD61+,GLY+
		Colonies CFC*	-

BM = Bone Marrow; PB = Peripheral Blood

* CFC: Colony Forming Cells

§ PBMC and BM MNC: Peripheral Blood-derived Mononuclear Cells and Bone Marrow-derived Mononuclear Cells respectively