Development of evolutionarily conserved viral integration sites as safe harbors for human gene therapy

Marco A. Quezada-Ramírez[1], Shannon Loncar[1], Matthew A. Campbell[2,3†], Krishna M. Parsi[1], Robert J. Gifford[4], and Robert M. Kotin[1*]

[1] Horae Gene Therapy Center
University of Massachusetts, Medical School
368 Plantation Street, Worcester MA 01605, USA

[2] University of Alaska Museum of the North
University of Alaska Fairbanks
1962 Yukon Dr. Fairbanks, AK 99775, USA

[3] Department of Animal Science
University of California, Davis
One Shields Ave, Davis, CA 95616, USA

[4] MRC-University of Glasgow Centre for Virus Research,
Bearsden, Glasgow, UK

*Correspondence to Robert.Kotin@umassmed.edu

[†] Present Address: School of Life and Environmental Sciences, The University of Sydney, 380 Werombi Road, Camden NSW 2570, AUS

Abstract

Gene transfer into CD34$^+$ hematopoietic stem and progenitor cells (HSPCs) involving integrating viral vectors has unpredictable outcomes including potential adverse events like leukemogenesis, resulting from insertional mutagenesis. Therefore, identifying and characterizing genome safe harbor (GSH) sites where exogenous gene material can be safely integrated into adult progenitor and stem cells is critically important for therapeutic gene addition. Here, we present a novel approach to identify new GSH sites based on a proven system of stable transgene insertion: the evolutionarily conserved integration of parvovirus DNA into the germlines of host species. From a dataset of 199 unique endogenous parvovirus (EPV) integration events identified in host species genomes, 102 loci were mapped to the human genome with 17 being experimentally evaluated as GSHs in primary human CD34$^+$ HSPCs. Nine promising GSHs resulted in cells edited using nucleofection alone or in combination with rAAV transduction. Of the nine GSH sites, six edited loci displayed sustained transgene expression in both erythroid and immune phenotypes while three clearly exhibited immune branch specific-regulation. Following this approach, additional GSH sites are likely to emerge from the remaining mapped loci for gene addition in hematopoietic stem and progenitor cells. Since it is unlikely that the GSH-lineage-restricted transgene expression is exclusive to hematopoietic stem cells, this approach extends the options for gene knock-ins while reducing the risks of insertional mutagenesis, unpredictable expression profiles, effects on differentiation, and increasing therapeutic effects.

Introduction

Genome safe harbors (GSH) may be defined at a high-level, as sites within the human genome where new genetic information can be introduced without disrupting normal cellular functions. In contrast, gene therapy applications that utilize integrating viral vectors, e.g., lentivirus or gammaretrovirus, tend to generate insertions randomly into transcriptionally more active and accessible euchromatin regions. Indiscriminate integration results in heterogeneous phenotypes of the transduced cell population with recombination events that can result in genotoxicity, gene silencing, unpredictable transgene expression or stability, loss of '"stemness", or dysregulated cellular proliferation, i.e., leukemias [1-3]. Thus, identifying GSHs that provide reliable locations for gene addition becomes a high priority for gene and cellular therapies. Clinical outcomes of lentiviral vector-dependent gene therapies addressing blood disorders (e.g. multiple myeloma [4], β-hemoglobinopathies[5], coagulopathies[6], Fanconi anemia[7], etc.) would improve upon selection of an appropriate GSH from an available catalog of verified GSH sites that reduce the risks associated with cell engineering. Importantly, the accessibility to primary human $CD34^+$ hematopoietic stem and progenitor cells (HSPCs) combined with programmable endonuclease technologies (particularly CRISPR/Cas9) constitute clinically relevant model to assess candidate GSHs for autologous cell therapies.

To date, efforts are ongoing to find new human GSH sites beyond the widely applied *AAVS1*[8-10], *CCR5*[11,12], and *hRosa26*[13-15] loci. Criteria have been proposed for selecting potential GSH sites that are intergenic and preferentially 50-300 kb distant from known or predicted transcriptional units[16-19]. However, the number of characterized

sites following this approach have been limited and mostly evaluated in established cell lines with few examples using primary human cells[20,21]. Other criteria proposed for identifying GSH sites are based on using endogenous mobile elements[22] or higher-order, chromatin architecture analysis[21,23] but outcomes have not been particularly effective. Thus, *AAVS1* which was developed as a GSH [10,15], remains a popular and widely used locus for directed transgenesis given that its knock-in does not affect the cell viability or differentiation [24,25] although some reports argue that transgene silencing occurs in a cell lineage[26] or promoter-specific manner[27]. The *AAVS1* site was originally identified as an AAV provirus integration site in exon 1 of *PPP1R12C* gene on human chromosome 19, within a motif consisting of AAV Rep protein-dependent origin of DNA synthesis [8,9,28], however, commonly available editing reagents relocated the site to the first intron of *PPP1R12C*[29]. As *AAVS1* is a parvovirus integration site, we hypothesized that other parvoviral integration loci may predict GSH sites too.

Parvoviruses are among the small number of viruses known to have contributed genes to the vertebrate germline via viral integration and horizontal gene transfer[30]. As a result, endogenous parvoviral elements (EPVs) have resulted from ancient infections across phylogenetically diverse host species that integrated into the germline and persisted during multiple speciation events across geologic time scale[31]. An extensive phylogenetic analysis of EPVs in vertebrate genomes disclose clear homology to members of extant parvovirus genera, including *Amdoparvovirus* [32], *Protoparvovirus* [33,34], and *Dependoparvovirus* [31,35]. The viral DNA endogenization represents an alternative to random mutation where large-step mutations resulted in the acquisition of novel viral “alleles,” facilitated genomic rearrangements, disrupted gene expression,

created novel transcription factor binding sites, or resulted in “exaptation” of viral genes with potential benefits to the host organisms[36], notably syncytin - a retrovirus envelope protein co-opted by placental mammals for trophoblast fusion early in embryogenesis[37]. Thus, identifying human orthologous sites to EPVs in non-human vertebrate genomes may indicate suitable sites for gene addition.

In this study, we identify 102 potential human GSH sites that are orthologs to EPV integration sites in the respective host species[38]. Orthologous loci were broadly categorized as intergenic or intronic according to the relative location in the host species and in the human genome. Additionally, the accessibility for genome editing in hematopoietic precursors was predicted by analyzing available ATAC-seq data. Through CRISPR/Cas9-mediated genome editing, seventeen of these loci (8 intergenic and 9 intronic) were evaluated as GSH sites by inserting an *eGFP* expression cassette in primary human $CD34^+$ HSPCs. The edited cells displayed no alterations in self-renewal or multipotency, i.e., stemness, demonstrating suitability of these loci for gene addition. Moreover, transgene expression was maintained across differentiation stages, indicating stable integration. Remarkably some GSHs displayed immune cell restrictive regulation. To our knowledge, this is the first study that provides an extended catalog of human GSH sites with potential applications for $CD34^+$ HSPC-based therapies. Furthermore, it is likely that lineage restrictive expression is not exclusive to HSPCs and can be extended to other tissues with additional GSH sites that are likely to emerge from the remaining mapped loci.

Results

Homology ascertainment relative to human genome

The local genomic landscape surrounding EPV insertions may be broadly categorized as intergenic or intronic. This categorization carries over when identifying the equivalent orthologous position in the human genome. For the 102 mammalian EPV loci (62 intergenic and 40 intronic) we mapped the orthologous locus in the human genome by comparative genomics, regional genome alignments and gene synteny[39], as well as similarity of DNA flanking EPVs using BLAT or BLASTN analysis. Relative positions between host species and human genome were established by interspecific collinearity and were refined to more precise coordinates based on sequence homology.

The putative GSH loci identified via this approach were found distributed across the human genome without obvious sequence or higher order common features (Fig. 1a; Supplementary Table 1). The majority of the EPVs were derived from members of the *Dependoparvovirus* and *Protoparvovirus* genera (Fig. 1b). An EPV loci nomenclature referencing the parvoviral genus from which each EPV is derived was established and can be consulted together with coordinates of their relative position on the human genome in the Supplementary Tables 1 and 2.

Selection of candidate GSH sites

Human ortholog loci from the intronic and intergenic categories were selected for experimental characterization (Supplementary Table 2). The proximity to open chromatin regions was evaluated using available ATAC-seq data for human $CD34^+$

HSPCs[40], with sites within 2 kb from an ATAC-seq peak prioritized for evaluation. Seventeen loci, 8 intergenic and 9 intronic, were selected for the first round of screening in HEK293T cells (Supplementary Table 2). Initially, single guide RNAs (hereafter gRNAs) were designed using online gRNA predictive tools (CHOPCHOP, CRISPOR, and CRISTA). Then, three to five highly scored gRNAs per locus (with minimal off-target prediction) were selected and incorporated into pX330 plasmids to co-express SpCas9 and the corresponding gRNA. HEK293T cells were independently transfected to select the best gRNA per locus using a highly cited *AAVS1* gRNA for comparison purposes[41] (Fig. 1c, Supplementary Table 3). Relative to *AAVS1* indel formation the GSH candidate loci indel frequencies suggested reasonable propensity to genome editing (Fig. 1c), therefore, we proceeded with a proof-of-concept study editing primary human $CD34^+$ HSPCs derived from bone marrow of healthy adult donors.

Editing of GSH candidates in human $CD34^+$ HSPCs with CRISPR/Cas9 and ssDNA templates

Several conformers of DNA templates have been reported for targeted integration into hematopoietic cells. These include reporter or therapeutic genes in plasmids or linear double stranded DNA (dsDNA) flanked by homology arms that mediate the homologous recombination (HR) [20,42], or flanked with *gRNA* recognition sites for homology-independent targeted integration (HITI) [43]. However, these dsDNA templates are transcriptionally competent as episomal DNA and elicit robust innate immune responses that may result in cytotoxicity. On the other hand, recombinant AAV6 vectors have been developed for HR and HITI but even when these vectors enhance the genome editing

outcomes [44], the DNA tends to remain episomal after the transduction and positive or negative strands that may anneal into transcriptionally active duplex DNA potentially leading to false positive editing results. Therefore, we decided to test single-stranded DNA (ssDNA) templates as these reportedly correct point mutations associated with blood disorders [45,46]. Our preliminary comparison of dsDNA and ssDNA templates displayed reduced GFP background with ssDNA templates (Supplementary Fig. 1A and B), reducing the random integration effects of episomal DNA templates following transfection (Supplementary Fig. 1C and D). Consequently, we engineered ssDNA templates per locus composed of an expression cassette consisting of *eGFP* open reading frame regulated by the MND synthetic promoter and rabbit beta-globin polyadenylation signal. The expression cassette was flanked by locus specific 300 nt homology arms yielding approximately 2 kb templates (Fig. 2a).

Thus, synthetic gRNAs complexed with SpCas9 were co-delivered together with 3 µg of the corresponding ssDNA template into $2.5x10^5$ primary human $CD34^+$ HSPCs following a nucleofection protocol (Lonza Inc., Morristown, NJ, USA). To minimize transcriptional interference, intronic GSH loci templates were engineered to insert the transgene in opposite orientation relative to the transcriptional unit hosting the candidate GSH site. Two days after nucleofection, the GFP^+ cells were collected through FACS sorting and cultured in SFEM liquid medium for additional 3 days (Fig. 2b). Unless noted differently, the experiments with undifferentiated $CD34^+$ HSPCs spanned seven days to avoid impairing the stemness features that occur following prolonged cell culture [47]. The GFP^+ percentages ranged from 0.12% ± 0.05 GFP^+ cells (mean ± s.d. $P < 0.0001$), for the intergenic locus *Dep.36*, up to 3.23 % ± 0.86 GFP^+ cells ($P = 0.8938$) for the intronic

locus *Dep.2* (Fig. 2c). The reference control for these experiments, locus *AAVS1*, reached 2.65% ± 0.88 GFP^+ cells (Fig. 2c). Not surprisingly, the GFP^+ percentages in $CD34^+$ HSPCs differed substantially from the values estimated through indels in human embryonic kidney (HEK) 293T cells that were typically ≥ 10% (Fig. 1c). Multiple factors account for the different outcomes including the HEK293T-suppressed innate immune responses, overall robustness of this cell line, and hetero-/euchromatization differences between cell types. For instance, the candidate GSH site *Dep.36* (intergenic) displayed high propensity to genome editing in HEK293T with approximately 30% indels formed (Fig. 1c), but yielded the lowest percentage of GFP^+ HSPCs after the nucleofection (Fig. 2c). Although we consulted available ATAC-seq data for $CD34^+$ HSPCs[40] as an additional criterion to select the best GSH candidates, no evident correlation was observed between editing efficiencies and the proximity to reported open chromatin spots (Supplementary Tables 2 and 3). An extreme example, is the *Dep.2* intronic locus that resulted in the highest percentage of GFP^+ cells at two days post-nucleofection (3.23% ± 0.86, P = 0.8938) despite having 15 kb distance from the nearest ATAC peak (Supplementary Table 3). Furthermore, persistent *eGFP* expression was monitored through differentiated stages of edited *Dep.2* cells, a phenomenon replicated by eight other loci lacking proximity to reported ATACseq peaks, namely the intergenic loci *Dep.3*, *Dep.22*, *Ap.102*, and the intronic sites *Dep.55*, *Prot.218*, *Prot.2*, *Prot.181* and *Dep.1* (Supplementary Table 3; Fig.2c and Fig. 4a).

The cell viability fluctuated between 75.15% ± 20 live cells (mean ± s.d., P = 0.0551), for the targeted *Prot.181* site, and 94.65% ± 3.71 live cells (P = 0.9994) for the targeted *Dep.13* site, as compared to the 98.4% ± 0.92 of the mock control (i.e.,

electroporated cells in absence of any CRISPR/Cas9-ssDNA editing set) (Fig. 2d). The proliferation rate decreased significantly within the three days following the sorting of GFP^+ HSPCs compared to mock control, 6.19-fold ± 1.96, excepting for *AAVS1* (4.43-fold ± 1.22, $P = 0.3294$), *Dep.1* (4.56-fold ± 1.53, $P = 0.4331$), *Dep.2* (4.21-fold ± 0.50, $P = 0.1373$), and *Dep.35* groups (5.17-fold ± 0.18, $P = 0.9885$) (Fig. 2e). Despite this, large colonies of differentiated cells formed independently of the edited GSH locus demonstrating that neither genome modification nor *eGFP* transgene expression impaired the proliferation capacity of the cells (Fig. 4a).

Targeting Long Term-HSCs $CD34^+$ populations

Primary $CD34^+$ HSPCs are a heterogeneous cell population whose developmental stages are identifiable by specific cell-surface markers. Within this, the long-term hematopoietic stem cells (LT-HSCs) are those with self-renewal and long term engraftment potential considered the most clinically relevant for autologous therapies [48]. In bone marrow, HSPCs are very scarce, estimated at 0.01% the cell population and LT-HSC in bone marrow has been estimated to be even scarcer, approximately 0.0001% of the cell population. To determine whether GSHs were targeted in LT-HSCs and cells survived gene addition, we analyzed the GFP^+ cell fraction to distinguish potential LT-HSCs ($CD34^+/CD90^+$) from the most abundant progenitor cells ($CD34^+/CD90^-/CD38^+$, $CD34^+/CD90^-/CD38^-$) [49]. Interestingly, the results suggest that not all loci are accessible in potential LT-HSPC populations (Fig. 2f). The loci that produced edited LT-HSPCs include *AAVS1* (10.32%), *Ap.102* (35.97%), *Dep.3* (12.07%), *Dep.28* (2.05%), *Dep.22* (8.56%), *Prot.181* (10.59%), *Dep.2* (8.33%), *Prot.176* (16.67%), *Dep.1*

(34.7%), *Prot.2* (11.18%), *Dep.35* (13.22%), *Dep.55* (26.09%), and *Prot.218* (5.6%) (Fig. 2f), making these attractive candidates for long term animal studies.

GSH targeted integration in human CD34$^+$ HSPCs

To verify template integration into the GSH candidate loci, genomic DNA was extracted from GFP$^+$ HSPCs sorted after 14 days of liquid cell culture. The DNA was analyzed by droplet digital PCR (ddPCR) with a locus-specific primer/probe set spanning the junction region and a second primer/probe set specific for an unedited part of the locus to obtain the total number of alleles (edited and unedited) in the sample (Fig. 2g; Supplementary Fig. 2; Supplementary Table 7). Considering two GSH alleles per diploid genome, a 50% ddPCR signal represents one hundred percent of monoallelic integration whereas >50% is indicative of partial bi-allelic integration. Thus, the results indicate that the editing of *AAVS1* (reference locus) occurred with 40% of bi-allelic integration (Fig. 2g), whereas the targeting of intergenic (*Dep.3*, *Dep.13*, *Dep.33*, and *Ap.102*) and intronic candidates (*Dep.1*, *Dep.2*, *Dep.35*, *Prot.2*, *Prot.181*, *Prot.176*, and *Prot.218*) resulted mostly in monoallelic integration likely resulting from more restricted accessibility (Fig. 2g). Notwithstanding, the integration of the transgene was stable for at least 45 days as demonstrated by the long-term culture of *AAVS1*, *Dep.2* (intronic), and *Dep.3* (intergenic) edited cells (Fig 2h). On the other hand, the relatively low level of GFP$^+$ cells obtained from the targeting of *Dep.34*, *Dep.36*, *Dep.28*, and *Dep.56* loci (Fig. 1c) resulted in difficult and sometimes undetectable integration events. Factors accounting for this outcome might include inaccessibility to the editing components, unfavorable ssDNA template secondary structure (e.g. the donor DNA template for the intronic *Dep.56* site displayed energetically stable secondary conformations), disruption of vital

pathways, or even HUSH-mediated repression of intronless transgenes [50], making them poor candidates for future *in vivo* transplantation studies.

Tracking studies of corrected HSPCs in Wiscott-Aldrich syndrome or β-hemoglobinopathy patients have estimated that as few as 98 corrected HSPCs per 10^6 infused $CD34^+$ cells are able to engraft the bone marrow for long term-production of granulocytes[48]. Given that our experiments were designed using relatively small cell numbers ($2.5x10^5$ to $1x10^6$ $CD34^+$ HSPCs per single nucleofection targeting one candidate GSH site), the results suggest that the scaling-up to clinical preparation numbers (usually $>5X10^6$ electroporated cells) would increase the bioavailability of therapeutic cells [44,51]. Alternatively, if low cell numbers are involved, the rAAV6 vectors reportedly substantially increase editing efficiencies. To test the latter approach, we combined nucleofection and rAAV6 transduction to deliver homology arm-carrying templates targeting three intergenic (*Ap.102*, *Dep.33* and *Dep.13*), and two intronic GSH loci (*Dep1* and *Dep.35*) which showed low efficiencies following ssDNA template nucleofection (Fig. 2c and Fig. 2i). Equivalent cell numbers were used (i.e., $2.5x10^5$ cells). As expected, at two days post transduction, the percentage of sorted GFP^+ cells increased substantially without affecting cell viability (Fig. 2i), but given that AAV transgenes persist episomal after internalization, accurately determining the integration level on sorted cells requires further analysis (Fig, 2h and Supplementary Fig. 1E). Regardless, these results demonstrate the feasibility to increase the targeting of apparently difficult GSH candidate loci.

Transcriptional changes in edited $CD34^+$ HSPCs

To determine whether and to what extent GSH editing perturbs the transcriptional homeostasis of stem and progenitor cells, RNAseq was performed with the bulk of sorted GFP^+ HSPCs from three intronic (*Dep.2*, *Dep.55* and *Prot.218*) and two intergenic (*Dep.3* and *Ap.102*) GSH loci targeted with ssDNA templates. Differentially expressed genes (DEGs), were determined by comparing edited versus unedited control cells from the same human donor (mock nucleofection). Consequently, the targeting of the reference GSH site, *AAVS1,* resulted in 1,595 DEGs, whereas the intronic GSH candidates *Dep.2*, *Dep.55*, and *Prot.218*, resulted in 2,163, 925 and 1,144 DEGs, respectively (Figure 3a and 3b). Perhaps unsurprisingly, the genome editing of the intergenic GSH sites *Dep.3* and *Ap.102* resulted in fewer DEGs reporting 509 and 496, respectively (Fig. 3a and 3b). Thus, even though the $CD34^+$ HSPCs constitute a heterogeneous population the results suggest that targeting intergenic GSH sites tends to perturb the transcriptome to a lesser extent than intronic GSH loci.

In addition, no substantial transcriptional changes on the genes proximal to the insertion sites were detected (Supplementary Fig. 3). Uniquely, the host gene for the intronic *Prot.218* locus (*ENPP2* gene) displayed a modest 2.95-fold increase on gene expression ($P = 2.4 \times 10^{-5}$) whereas a downstream transcript (*DEPTOR*) decreased by 0.18-fold ($P = 2 \times 10^{-11}$) (Fig. 3d and Supplementary Fig. 3). Similarly, transcripts mapping to *GPNMB*, downstream to *Dep.3* locus, displayed a fairly robust, 0.34-fold decrease relative to unedited control cells ($P = 0.03$) (Fig. 3d and Supplementary Fig. 3).

The gene ontology (GO) analysis of DEGs ranked more than one group of genes associated to regulation of cell proliferation (GO:0008285, GO:0042127,

GO:0001558, GO:0008284, and GO:0051726). Notably, the genes associated with malignant transformation showed no evidence of alteration (Fig. 3c). The evaluated transcriptomes revealed upregulation of tumor suppressor genes *CDKN1A* (8.52-fold change, $P \leq 10^{-8}$), *INHBA* (29.1-fold change, $P \leq 0.001$), and *TP53I11* (2.5-fold change, $P \leq 0.01$) (Figs. 3b and c). In contrast, RNAseq analysis showed downregulation of leukemia-associated proto-oncogenes such as *CCND1/2* (0.46-fold change, $P < 0.05$), *CCNE2* (0.38-fold $P = 0.0076$), *MYC* (0.42-fold change, $P < 0.001$), *SRC* (0.3-fold change, $P < 10^{-5}$), or *MPL* (0.25-fold change, $P < 0.05$) (Fig. 3c). Additionally, intergenic GSH sites *Dep.3* and *Ap.102* displayed downregulation of the AML-associated proto-oncogene *FLT3* (0.44-fold change, $P \leq 0.0014$) and overexpression of the tumor suppressor gene *INHBA* (18.4-fold change, $P \leq 0.005$) (Fig. 3c). With the exception of the reference *AAVS1* edited cells, all the GSH-targeted cells underwent overexpression of *HEY1* (98.8-fold, $P < 10^{-40}$, for *Dep.3*, *Dep.55*, *Ap102*, and *Prot,218*), or upregulation of *HES7* (8.9-fold $P = 0.017$ for *Dep.2*), both transcriptional repressors for the maintenance of blood precursors (Fig 3d). In addition, overexpression of *FRZB* (74.2-fold change, $P < 0.01$), an inhibitor of the □-catenin pathway, was observed (Fig. 3b and 3d). Similarly, all the edited GSH groups displayed up-regulation of genes associated with innate immune response presumably resulting from *in vitro* manipulation, i.e., nucleofection of ssDNA, CRISPR:Cas9 complex, etc. (Fig. 3c). Together these changes confirmed the observed phenotype of the manipulated $CD34^+$ HSPCs in culture (i.e., reduced cell proliferation, Fig. 2e) and indicate low propensity for malignant transformation, however, deeper sequencing (with homogeneous stem cell

systems) as well as cell transformation assays will help to clarify the cellular response to and assess risks associated with editing of the novel GSH sites.

Edited CD34$^+$ HSPCs retained the multipotency with widespread or lineage-specific transgene expression

To determine whether gene addition and transgene expression affected the HSPC proliferation or differentiation, cells were examined following 14 days of culture in differentiation medium. No morphological abnormalities were detected among the colonies (CFUs) derived from manipulated HSPCs (Fig. 4a). Instead, differentiated edited cells formed colonies that continued to express *eGFP*. *Dep.2*, *Dep.3*, *Prot.218*, *Prot.2* and *Prot.181* edited cells produced GFP$^+$ colonies of four lineages, erythroid (BFU-E), granulocyte (CFU-G), granulocyte-macrophage (CFU-GM) and macrophage (CFU-M) as was done for the reference group of edited *AAVS1* cells (Fig. 4a). In contrast, although *Dep.35* and *Proto.176*-derived colonies also showed widespread expression across the four lineages, the GFP$^+$ signal was dramatically reduced upon differentiation (Fig. 2g and Fig.4a) indicating downregulation of *eGFP* from these loci. Other edited loci resulted in *eGFP* expression biased towards granulocyte/macrophage lineages based on visual inspection of the colonies (*Dep.13*, *Dep.1*, *Dep.55*, *Ap.102* and *Prot.176*) (Fig.4a). Despite dissimilar expression phenotypes, the total colony distributions were similar among all the groups as compared to unedited control cells (Fig. 4b). Interestingly, the edited GSH cells with higher GFP signal and widespread *eGFP* expression by 14 days of differentiation were the reference locus, *AAVS1*, and *Dep.3*, *Dep.2*, *Dep.55*, and *Prot.218* loci (Fig. 4a and 4c) being included as well among

347 those loci resulting in targeting of LT-HSCs (Fig. 2f). These characteristics are attractive
348 for gene addition in human CD34$^+$ HSPCs.

349 To confirm the multipotency of edited HSPCs, we analyzed the erythroid
350 (CD235a$^+$) and immune cell (with the general CD69$^+$ cell surface marker) composition of
351 the GFP$^+$ differentiated population (Fig. 4c and 4d). The resulting compositions were
352 similar among most analyzed groups relative to the *AAVS1* edited reference locus (Fig.
353 4d and 4e). Similar percentages of erythroid and immune cell phenotypes coexisted
354 with small portions of unidentified GFP$^+$ or GFP$^+$/CD235a$^+$/CD69$^+$ triple labeled cells
355 (Fig. 4e). Remarkably, the bias towards non-erythrocyte cell expression from *Dep.1*,
356 *Dep.55*, and *Ap.102* GSHs was confirmed within the CD69$^+$ fraction (Fig. 4e).

357 To characterize further the immune specific *eGFP* regulation, we established
358 K562 cell-line models of *eGFP* integrated into *Dep.1*, *Dep.55*, *Dep.33,* or *Ap.102* GSHs.
359 The K562 cell lines were then induced to erythroid (CD235a+) or megakaryocytic
360 phenotype (CD61+) by treating with erythropoietin (hEPO) or phorbol myristate acetate
361 (PMA), respectively (Supplementary Fig. 5a). K562-derived megakaryocytes may be
362 considered as having an immune cell phenotype due to the recently described immune
363 role of megakaryocytes[52] and their separated origin from erythrocytes[53]. Thus, following
364 induction with PMA, the megakaryocytic cells retained strong expression of the *eGFP*
365 as expected (Supplementary Fig.5c). To determine whether epigenetic modifications
366 were associated with transgene expression, CUT&RUN assays coupled with qPCR
367 were performed. The results demonstrated that histone modifications showed
368 enrichment of H3K4me3, but not H3K27me3, associated with the MND promoter of the
369 transgene inserted in *Dep.1*, *Dep.55*, *Ap.102*, or *Dep.33* loci. Histones with H3K4me3

modifications are consistent with a sustained expression within immune cell lineages (Fig. 4d and Supplementary Figs. 5b, 5c and 6). In contrast to megakaryocytic induction, edited K562 cells treated with hEPO resulted in approximately 30% of induced cells positive for CD235a, an erythroid cell surface marker(Supplementary Fig.5d). Furthermore, the CD235a^{+} cell population was comprised of both GFP+ or GFP- cells. Similar results were obtained with erythroid induction with hemin (data not shown). DNA analysis confirmed the presence of properly integrated *eGFP* gene, thus GFP expression was most likely affected by epigenetic modifications; sequence analysis of the *eGFP* cassette would substantiate non-genetic basis for the GFP silencing (Supplementary Fig. 5d). These results are distinct from primary human HSCs in which the induced lineages were phenotypically homogeneous. Considering the leukemic origin of K562 cells, the model does not preclude the existence of immune specific GSHs observed after differentiation of modified CD34+ HSPCs.

Persistent transgene expression in terminally differentiated erythroid cells

As the final step of erythroid differentiation comprises the erythroblast enucleation, we selected *Dep.2*, *Dep.3*, *Prot.218*, *Prot.2* and *Prot.181* groups for further erythrocyte maturation experiments with hEPO (Fig. 4g). Following the nucleofection of CD34^{+} HSPCs to target the mentioned GSH sites, the cells were expanded during 14 days in culture under erythroblast stimulating conditions (Fig. 2b). Next, the erythroblasts were exposed to hEPO for an additional five days of culture after which enucleation was confirmed in GFP^{+} cells, but not in non-hEPO treated cells (Fig. 4h). Of note, persistent

transgene expression was observed in reticulocytes derived from all analyzed loci except for *Prot.218* (Fig. 4h) in which the transgene expression was silenced soon after exposure to hEPO (Fig. 4f).

Overall, the results support our approach for identifying useful GSH sites compatible with stable transgene expression, and in some cases tissue specific-expression helping to establish a concise catalog of sites for gene therapy of blood disorders.

Discussion

This study demonstrated the feasibility of using evolutionary biology and comparative genomic approaches to identify prospective GSH in the human genome. Over the course of millions of years, parvovirus DNA genomes have integrated into the germline of multiple host species at genomic loci that tolerate DNA insertions. Of the previously 200 unique EPV insertions identified [38], we have mapped about half to orthologous sites in the human genome indicating that these loci are broadly conserved evolutionarily. Genomic loci were selected for $CD34^+$ HSPC genome editing based on bioinformatic analyses which (i) determined precision of the EPV ortholog in the human genome (Fig. 1a, 1b, and Supplementary table 1); (ii) predicted editability, and (iii) predicted transcriptional activity in human cells. In contrast to alternative approaches for identifying GSH candidate loci that rely on arbitrary assumptions considered to reduce risks of insertional mutagenesis [16,19,21], the methodology applied in the current study was intended to identify and characterize hematopoietic-compatible GSH loci. The criteria for establishing an ontogenic GSH (e.g., as in germline editing for transgenic

mouse generation) are much different than the criteria used to establish a cell-, or lineage-, specific safe harbor, and as such, the “global” GSH is likely to be suboptimal for adult stem cells.

Following our approach, we evaluated and characterized novel genomic loci for gene therapy of blood disorders using primary human $CD34^+$ HSPCs. However, further evaluation with *in vivo* models is necessary to demonstrate the ability of modified HSPCs for long-term engraftment of the bone marrow and reconstitution of blood cell compartments, ideally using a non-human primate model. As a first approximation, the *in vitro* results presented here indicate successful editing among the rare LT-HSC population ($CD34^+$,$CD90^+$) [49] even with relatively low numbers of $CD34^+$ HSPCs used for editing. For these experiments, the number of nucleofected cells ranged from $2.5x10^5$ to $1.0x10^6$ for each GSH site being evaluated. In comparison, approved gene therapies against ADA-SCID (Strimvelis™, Orchard Therapeutics, PLC), MLD (Libmeldy™, Orchard Therapeutics, PLC) or β-thalassemia (ZYNTEGLO™, betibeglogene autotemcel, Bluebird Bio, Inc.) reportedly require up to $2.0\text{-}3.0x10^7$ autologous $CD34^+$ HSPCs per kg of body weight, retroviral or lentivirally transduced, as the dose for infusion. However, indiscriminate lentivirus integration generates a heterogenous population of cells including cells that are killed outright, cells that lose stemness, and cells that do not contribute any therapeutic benefit. Therefore, logic supports the rationale that targeting an appropriate GSH with sufficient editing efficiency would yield a population of productive stem cells or CAR-T cells, and would reduce the number of apheresis rounds to obtain an appropriate number of cells for manufacturing.

Contrary to the large number of chimeric antigen receptor (CAR) T-cells reportedly required in cancer clinical trials (up to $1.0x10^8$ cells/kg) [4,21,54], expanding human $CD34^+$ HSPCs by prolonged cell culture (>7 days) has been reported to detrimentally affect the long-term progenitor cells repopulating capacity [47]. Recent studies of long-term HSPCs transplanted in patients with Wiskott-Aldrich syndrome or β-hemoglobinopathies estimated that as few as 98 edited HSPCs per 10^6 infused $CD34^+$ cells correspond to engrafting cells were sufficient to reconstitute gene-corrected granulocytes in circulation [48]. This suggests that the more efficient the editing pipeline (which is partially dependent on efficient access to the genomic site), the greater the bioavailability of therapeutic cells in the infusion. If the novel GSHs are comparable to the highly efficient *AAVS1*, the up-scaling for clinical readiness (usually $>5x10^6$ electroporated cells) might result in increased numbers of engraftable cells (i.e., modified LT-HSCs).

Although the GFP^+ percentages of sorted cells were low for some targeted GSH loci (e.g. *Dep.1* or *Ap.102*), the values might easily be enhanced by cell cycle (S/G2) synchronization, screening additional gRNAs, or combined nucleofection/rAAV6 transduction as was demonstrated for *Ap.102*, *Dep.1*, *Dep.33* and *Dep.35* (Fig. 2h). In addition, persistent transgene expression in differentiated cells indicates that most of these GSH sites are compatible with long-term transgene expression making unnecessary the use of insulators to protect the transgene expression as discussed elsewhere [21,55,56]. In this regard, a naturally occurring insulator in *AAVS1* [57] might explain the robust expression reported from this site (Fig. 2c and 4d) [58] suggesting that insulators artificially incorporated into the transgene cassette may improve the

performance from targeted GSH sites. Whether the novel GSH sites contain insulators has not been determined, and the ssDNA templates used in this report were devoid of such elements. Nevertheless, constitutive reporter expression was maintained in immature as well as in both partially and terminally differentiated cells indicating that selected GSH loci are compatible with the long-term transgene expression in blood cells. Thus, these results establish a catalog of GSH with opportunities to engineer complex transgenes exploiting the permissiveness and tissue specificity of the novel GSH sites by including enhancers, insulators, IRES, T2A, or combinations that increase the transgene performance, or to engineer multigenic knock-ins that increase the traits of genetically modified hematopoietic cells. Remarkably, immune specific GSH sites introduce an additional level of safety restraining the transgene expression to the target cell population, particularly if intergenic GSHs are included as their modification conveys lower disturbances of the cellular transcriptome (Fig. 3a and 3b; Fig. 4). Similar tissue-specific GSHs are predicted to exist for all human tissue types.

In conclusion, this study established a concise catalog of 102 mapped GSH sites in the human genome with a sample of 17 loci yielding at least 9 GSH sites suitable for blood disorders therapy. Remarkably, the naturally occurring regulation from some GSH sites can be exploited for lineage restricted transgene expression. Thus, more GSH sites are likely to emerge from the remaining mapped loci with tissue specific-applications to modify other cell types.

Methods

Homology ascertainment relative to human genome

EPV positions were mapped to the orthologous human loci using Ensembl for comparative genomic alignments (GRCh38.p13) (Ensembl 2023). However, since the host species genome sequences accessible via Ensembl may be from different assembly versions than the reported EPV loci sequences the EPV coordinates within host genomes were acquired through BLASTN or BLAT analysis tool within the Ensembl suite. Then, using the Comparative Genomics options in Ensembl, the aligned host species and multiple mammalian genomes were analyzed to ascertain overall synteny and collinearity. Thus, the corresponding human genomic locations were determined. As some EPV host species were not available in Ensembl, BLASTN and EPV genomic flanking sequence data were used to locate the EPV integration site in closely related host species genomes that may or may not contain an EPV ortholog. Gaps in the host/human sequence alignment surrounding the EPV elements introduced another level of ambiguity in identifying sites in which the element may have been inserted.

Selection of human genome loci to be evaluated as GSH

We focused on EPV loci that were mapped with a high degree of confidence on the human genome. As the experimental portion of this effort involved editing human primary $CD34^+$ HSPCs, we evaluated the accessibility of chromatin for EPVs in this cell type. $CD34^+$ HSPC data were retrieved from the Gene Expression Omnibus under

accession GSE74912. ATACseq data were accessed with DolphinNext (Corces, et al.[39]) and chromatin accessibility was estimated through peak-calling with MACS2 with special attention for loci within 2 kb of an ATACseq peak.

Cell culture and lipofection

Human embryonic kidney (HEK)-293T cells (ATCC CRL-3216) were cultured in 10% FBS DMEM (Gibco 11965-092), supplemented with penicillin (50 U per mL) and streptomycin (0.1 mg per mL) (Sigma P4458), at 37°C in 5% CO2 and saturating humidity. Transfection experiments on HEK-293T cells were performed in 6-well plates (Corning, NY) . Briefly, one day prior to transfections, $3.5x10^5$ cells were added to each well. Cells were transfected with pX330 plasmid (2.5µg) using Lipofectamine 3000 (Invitrogen) following the manufacturer's instructions. Forty-eight hours post-transfection genomic DNA (gDNA) was isolated with DNeasy Blood & Tissue Kit (Qiagen), and following concentration measurement, an aliquot of gDNA was used for PCR amplification of the targeted loci using locus specific primers to determine editing efficiency and specificity.

Human bone marrow-derived $CD34^+$ stem and progenitor cells (HSPCs), obtained from healthy adult donors, were purchased from Stem Cell Technologies (SCT, Cambridge, MA 02142). Briefly, cryopreserved HSPCs were thawed and pre-stimulated for 48 h in StemSpanTM SFEM II medium (SCT 09655) with Expansion supplement (SCT 02691) and 1 µM UM729 (SCT 72332), and incubated at 37° C in 5% CO_2 and saturating humidity atmosphere. Following electroporation, the $CD34^+$ HSPCs

were returned to the same incubation conditions for additional 72 h before further downstream use.

Erythroleukemia K562 (ATCC CCL-243) cells were maintained in 10% Iscove's modified Dulbecco's medium(IMDM) (10-016-CV, Corning Life Sciences) supplemented with penicillin (50 U per mL) and streptomycin (0.1 mg per mL) (Sigma P4458), at 37° C, 5% CO2 and saturating humidity. Following electroporation, the cells were returned to the growth medium and incubated for 5 days before cell sorting. Sorted cells expressing GFP were expanded under the same incubation conditions for use in subsequent experiments.

Guide RNAs for CRISPR/Cas9 gene editing

Single guide RNAs (gRNAs) were simultaneously engineered with CHOPCHOP, CRISPOR and CRISTA designing tools. Three to five highly scored gRNAs with minimal off-target sites that were common to the three gRNA design programs, were chosen per candidate GSH locus. Synthetic oligonucleotides incorporating the *gRNA* sequences were ligated into pX330 plasmid (Addgene, Watertown, MA 02472) to co-express *Cas9* and *gRNA*. Indel formation of the target site was used as an indicator for editing efficiency in HEK-293 cells. Forty-eight hours post-transfection into HEK293T cells, genomic DNA was isolated (Qiagen), and PCR amplified into 600 bp products (Takara). Indel frequencies were determined through Sanger data-decomposition using the TIDE analysis tool (tide.nki.nl/). Indel efficiencies were compared to that provided for a widely reported *AAVS1* gRNA (Supplementary Table 3). Thus, the most efficient gRNA per locus was selected for downstream experiments in $CD34^+$ HSPCs. Next, a chemically

synthesized version of each gRNA, modified at both termini with 2'-O-methyl-3'-phosphorothioate, was purchased as a single guide molecule from IDT (Integrated DNA Technologies, IA) and complexed with ArchiTect™ Cas9 nuclease (Stem Cell Technologies) for editing experiments in primary human $CD34^+$ HSPCs.

Single-stranded DNA template preparations

Single-stranded DNA donor template (ssDNA) was synthesized from a plasmid carrying 300 bp homology arms matching the corresponding GSH site flanking the *eGFP* expression cassette. Briefly, an expression cassette containing the *eGFP* open reading frame (ORF) regulated by the MND promoter and rabbit β-globin poly(A) signal was constructed by using conventional molecular biological techniques. Homology arms were obtained by PCR from HEK-293T genomic DNA and a plasmid template was generated per candidate GSH site by Gibson assembly (New England Biolabs) using fragments synthesized by Q5 High Fidelity Taq polymerase (New England Biolabs). Finally, 5'-phosphorylated PCR primers were used to amplify the 2.0 kb eGFP expression cassettes flanked with 300 bp homology arms (Supplementary Table 6). To produce single-stranded template DNA, the phosphorylated reverse strand from each substrate was digested by sequential strandase-treatment to yield ssDNA (Takara 632666) which was column-purified and resuspended in nucleases-free water (Gibco). Plasmid templates were eliminated using DpnI treatment (NEB R0176) prior to the ssDNA synthesis..

Electroporation

571 Prior to nucleofection of $CD34^+$ HSPCs, ArchiTect™ Cas9 nuclease (SCT) and the
572 corresponding gRNA were combined (1:2.5 molar ratio) to form a ribonucleoprotein
573 complex at 25° C for 20 min and placed at 4° C until used for nucleofection reactions. In
574 parallel, $CD34^+$ HSPCs pre-stimulated for two days were washed in 50 mL PBS at 37°
575 C and resuspended in P3 nucleofector solution at RT (Lonza, V4XP-3032).
576 Approximately $2.5x10^5$ $CD34^+$ HSPCs were mixed with gRNA:Cas9 complexes and 3
577 µg ssDNA donor template (20 µL) in nucleofector strip format (Lonza). Electroporation
578 was performed in Unit X, Lonza 4D nucleofector, using the DZ-100 program. After
579 nucleofection, cells were incubated for 10 min at room temperature before adding the
580 StemSpan™ SFEM II medium with Expansion supplement and 1 µM UM729 for a 2
581 day-recovery time. After the recovery period, the cells were sorted by flow cytometry
582 and GFP^+ cells were collected. Similarly, erythroleukemia K562 cells were washed in 50
583 mL PBS at RT and approximately $1x10^6$ cells were resuspended in SF solution (Lonza
584 V4XC-2032) and electroporated in presence of pre-assembled Cas9:gRNA complexes
585 plus 12 µg ssDNA (100 µL format) in a Lonza 4D-nucleofector using the FF-120
586 program.
587

588 ## AAV6 transduction

589 Recombinant adeno-associated virus type 6 (rAAV6) vectors were obtained from the
590 University of Massachusetts Chan Medical School Viral Vector Core facility.
591 Cryopreserved $CD34^+$ HSPCs were thawed, transferred to tissue culture plates, and
592 incubated in StemSpan™ SFEM II medium for a 2 days pre-stimulation period under
593 conditions described above. Following pre-stimulation treatment, $2.5x10^5$ $CD34^+$ cells

were electroporated in the presence of Cas9:gRNA complexes, followed by a recovery period of 15 min at RT. Next, cells were collected and transduced with rAAV6 at a multiplicity of infection (MOI) of 200,000-300,000 gc/cell (genome copies) in a 15 mL conical tube (Corning). Cells were transferred to a 6-well culture plate (Corning) at a density of $1x10^5$ cells/mL and incubated for 2 days to recover from the treatment before downstream analysis.

Colony Forming Unit assay (CFU)

Between 300 and 1500 sorted GFP^+ cells were seeded in 1 mL methylcellulose hematopoietic progenitor cell differentiation medium (MethoCult, SCT H4435) and transferred to 6-well SmartDish plates (SCT 27371). Differentiation was induced during 14 day incubation at 37°C, 5% CO_2 and saturating humidity before scoring the Colony Forming Units (CFUs) by brightfield microscopic inspection (Axiovert 135 microscope, Zeiss) and fluorescent images were obtained with fluorescent microscopy (LionHeart FX,BioTek). After colony scoring, the entire plate of cells were recovered from methylcellulose by diluting in PBS, washed and resuspended in ice cold 1% FBS in PBS before the staining for flow cytometry analysis as described above.

Flow cytometry

After recovering for 2 days, nucleofected $CD34^+$ HSPCs were washed once in PBS and resuspended in ice cold phenol red- free-IMDM (Gibco) supplemented with 1% FBS (Gibco) and processed using fluorescence activated cell sorting (FACS). Non-transfected HSPC samples were used for gating GFP^+ signals. Propidium iodide

(Sigma) was included for viability determinations then, GFP^+ cells were collected and cultured for 3 additional days in StemSpan™ SFEM II, $CD34^+$ Expansion supplement and 1 µM UM729. The fold-expansion was determined by dividing the number of cells after three days of culture by the number of GFP^+ cells seeded immediately after sorting. Thus, the total timeline for nucleofection experiments starting with cryopreserved, immature $CD34^+$ HSPCs spanned 7 days. For the analysis of immature populations (LT-HSC and MPPs), cells were cultured for five days post-nucleofection, washed in PBS and then resuspended in ice cold PBS supplemented with 1% FBS. Subsequently, cells were incubated in presence of TruStain FcX™ (BioLegend) before staining with antibodies against CD34 (BioLegend 343625), CD90 (BioLegend 328114) and CD38 (BioLegend 303506). Samples were fixed in ice cold 2% paraformaldehyde (PFA) for 10 min, washed and resuspended in phenol red free-IMDM. The $GFP^+/CD34^+$ population was gated to explore the modified immature cell populations. For cell lineage determinations (erythroid and immune), cells were recovered by diluting the methylcellulose in PBS and resuspending the cells in ice cold 1% FBS in PBS. The Fc receptors were blocked with Human TruStain FcX™ (BioLegend) before staining with antibodies against CD235a and CD69 for 20 min at 4°C in the dark. All antibodies used in this work can be consulted in Supplementary Table 4. Following the incubation, the cells were washed in ice cold PBS and resuspended in ice cold 2% PFA for 10 min, washed once again with ice cold PBS, and finally resuspended in 1% FBS phenol red-free IMDM (Gibco) before the analysis. For experiments with differentiated K562 cells, the cells were stained with antibodies against CD235a (erythroblastic) or CD61 (megakaryocytes) surface markers as described above except that fixation with 2% PFA

was omitted as the samples were processed immediately for CUT&RUN epigenetic modification analysis.

Erythroid differentiation

Nucleofected CD34$^+$ HSPCs were maintained in SFEM II cell culture medium with Erythroid supplement (SCT) to induce erythroblast expansion during 14 day incubation. Medium changes were performed at days 7 and 11 of culture. At day 14, cells were transferred into SFEM II supplemented with 3 U/mL hEPO (SCT) and 3% AB human serum (Sigma) for an additional 5 days of culture. Cell nuclei were stained with Hoechst 33342 (ThermoFisher) prior to microscopy imaging.

Erythroblastic and Megakaryocyte induction of K562 cells

K562 cells were induced to erythroblastic cells by exposure to 3U/mL hEPO (SCT 78007) or 50 ìM hemin (Sigma). Megakaryocytic cells were obtained with 80 nM phorbol 12-myristate 13-acetate (Invivogene). Cellular phenotypes were identified by cell surface markers using anti-CD235a antibody (BioLegend 349105) for erythroblastic and anti-CD61 antibody (BioLegend 336406) for megakaryocytic cells and flow cytometry. The sorted cells were immediately processed for CUT&RUN experiments.

Droplet digital PCR

The integration efficiency was determined by ddPCR of genomic DNA extracted from sorted GFP$^+$ cells of 14 days post-nucleofection. Briefly, genomic DNA isolated from the cells (Qiagen) was amplified by ddPCR in the presence of a primer/FAM probe set

specific for the integrated transgene and a primer/HEX probe set specific for a reference region within the same unaltered locus (Supplementary Tables 7, 8 and 9). Reaction mixtures consisted of 1x ddPCR Supermix for probes (No dUTP) (Bio-Rad), primer/probe mixtures (900 nM primer/250 nM probe) and up to 60 ng gDNA per reaction. The thermocycler was programmed as follows: initial denaturation (95°C for 10 min with ramp rate of 2°C/sec), followed by 60 cycles of denaturation (94°C for 30 sec, ramp rate of 2°C/sec), annealing (55°C for 1 min ramp rank of 1°C/sec), and extension (72°C for 2 min, ramp rate of 1°C/sec), and a terminal denaturation step (98°C for 10 min), followed by cooling (4°C). Droplet Reader and Quantasoft analysis software (Bio-Rad) were used to analyze the fluorescence using 1D and 2D plots for thresholding FAM and HEX positive droplets. Ratios of FAM to HEX were used to calculate the abundance of edited alleles.

RNA-seq

Sorted GFP^+ HSPCs were cultured for three days in StemSpanTM SFEM II medium (SCT) before being washed in PBS. Subsequently, the cells were mixed with five volumes of RNA*later* solution (Thermo Fisher AM7020) and stored at -80° C until processing for RNA-seq analysis. For library preparation, the mRNA fraction was prepared by poly(A)+ selection (Illumina). Libraries were sequenced as paired end 150 bp reads on an Illumina HiSeq 3000 with a sequencing depth of approximately 20-30 million per sample. Quality control metrics were confirmed in FASTQ files through FastQC (bioinformatics.babraham.ac.uk/projects/fastqc/). Adapter sequences were removed from the reads and mapped to the *Homo sapiens* reference genome

(GRCh38) available on ENSEMBL using STAR aligner v.2.5.2b. Differential expression analysis was performed by analyzing un-normalized read counts using DESeq2 and applying “donor” as a cofactor. Transcripts with an adjusted p-value < 0.05 and absolute $\log_2$ fold change > 1 were designated differentially expressed genes (DEGs) as d. Heatmaps with $\log_2$ normalized read counts were generated with R Studio, clustering genes according to their regulation changes. Gene ontology (GO) analysis was performed with the iPathwayGuide platform of Advaita Bioinformatics (advaitabio.com).

CUT&RUN Assay

Chromatin samples were processed using the EpiCypher CUTANA™ ChIC/CUT&RUN kit (EpiCypher 14-1048), according to the manufacturer’s protocol. Briefly, approximately 500,000 cells per sample were collected by centrifugation, resuspended in PBS, then re-centrifuged and then coupled to activated Concanavalin A beads. The cells were incubated with antibodies overnight at 4°C. Approximately 0.5ug of an antibody was used per sample, including IgG negative control antibody (EpiCypher 13-0042), H3K4me3 (EpiCypher 13-0041), and H3K27me3 (Active motif 39055). For permeabilization, 0.01% of digitonin was used in the buffers. Unbound antibody was removed by washing 2x, the cells were incubated with protein A and protein G - micrococcal nuclease fusion (pAG-MNase) to recover the histone - DNA complex. MNase was activated by adding 1mM CaCl2 and incubated at 4°C for 2 hours, releasing the chromatin immune complexes into the supernatant. The MNase reaction was terminated using a stop buffer master mix and to remove RNA from the sample. The released DNA fragments were recovered and size selected using SPRIselect beads and

eluted in 15ul of elution buffer for subsequent qPCR analysis. Genomic DNA extraction was extracted from the input sample (1% of total cells) and the enrichment of immunoprecipitated DNA was calculated using percent input method formula - 100*2^(Adjusted Input – CT (IP). The final values were plotted as fold enrichments over IgG negative control.

Statistical analysis

All the experiments with primary human cells were independently performed at least three times for statistical analysis. Cells were harvested from different anonymous, healthy BM donors each time. Data represents mean ± SD. Statistical analyses were performed by ANOVA with Dunnett test post hoc. Significant and non-significant *P* values are indicated on the Figures and the main text. Results were considered significant when $P < 0.05$. When less than 3 experiments were performed for descriptive results, an opportune mention is indicated on the figure legends.

Supplemental Information

Supplemental Documents SX – Locations of EPVs on Human Genome (eve-meta-hg38)

Data Availability

Sequence data generated for this study have been deposited in the National Center for Biotechnology Information Sequence Read Archive under BioProject PRJNA1022496.

References

1. FDA Investigating Serious Risk of T-cell Malignancy Following BCMA-Directed or CD19-Directed Autologous Chimeric Antigen Receptor (CAR) T cell Immunotherapies. FDA, U.S. Food & Drug Administration, www.fda.gov/vaccines-blood-biologics/safety-availability-biologics/fda-investigating-serious-risk-t-cell-malignancy-following-bcma-directed-or-cd19-directed-autologous (2023).
2. Strimvelis (autologous CD34+ enriched cell fraction that contains CD34+ cells transduced with retroviral vector that encodes for the human adenosine deaminase cDNA sequence): first case of lymphoid T cell leukaemia after insertional oncogenesis EMA, European Medicines Agency www.ema.europa.eu/en/medicines/dhpc/strimvelisr-autologous-cd34-enriched-cell-fraction-contains-cd34-cells-transduced-retroviral-vector (2021).
3. Jones, R. J. & DeBaun, M. R. Leukemia after gene therapy for sickle cell disease: insertional mutagenesis, busulfan, both, or neither. *Blood* 138, 942-947, doi:10.1182/blood.2021011488 (2021).
4. Stadtmauer, E. A. et al. CRISPR-engineered T cells in patients with refractory cancer. *Science* 367, doi:10.1126/science.aba7365 (2020).
5. Dever, D. P. et al. CRISPR/Cas9 beta-globin gene targeting in human haematopoietic stem cells. *Nature* 539, 384-389, doi:10.1038/nature20134 (2016).
6. Gong, J., Chung, T. H., Zheng, J., Zheng, H. & Chang, L. J. Transduction of modified factor VIII gene improves lentiviral gene therapy efficacy for hemophilia A. *J Biol Chem* 297, 101397, doi:10.1016/j.jbc.2021.101397 (2021).
7. Diez, B. et al. Therapeutic gene editing in CD34(+) hematopoietic progenitors from Fanconi anemia patients. *EMBO Mol Med* 9, 1574-1588, doi:10.15252/emmm.201707540 (2017).
8. Kotin, R. M., Linden, R. M. & Berns, K. I. Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination. *EMBO J* 11, 5071-5078, doi:10.1002/j.1460-2075.1992.tb05614.x (1992).
9. Kotin, R. M. et al. Site-specific integration by adeno-associated virus. *Proc Natl Acad Sci* U S A 87, 2211-2215, doi:10.1073/pnas.87.6.2211 (1990).
10. 1Smith, J. R. et al. Robust, persistent transgene expression in human embryonic stem cells is achieved with AAVS1-targeted integration. *Stem Cells* 26, 496-504, doi:10.1634/stemcells.2007-0039 (2008).
11. Gomez-Ospina, N. et al. Human genome-edited hematopoietic stem cells phenotypically correct Mucopolysaccharidosis type I. *Nat Commun* 10, 4045, doi:10.1038/s41467-019-11962-8 (2019).
12. Scharenberg, S. G. et al. Engineering monocyte/macrophage-specific glucocerebrosidase expression in human hematopoietic stem cells using genome editing. *Nat Commun* 11, 3327, doi:10.1038/s41467-020-17148-x (2020).

13. Irion, S. et al. Identification and targeting of the ROSA26 locus in human embryonic stem cells. *Nat Biotechnol* 25, 1477-1482, doi:10.1038/nbt1362 (2007).
14. Zambrowicz, B. P. et al. Disruption of overlapping transcripts in the ROSA beta geo 26 gene trap strain leads to widespread expression of beta-galactosidase in mouse embryos and hematopoietic cells. *Proc Natl Acad Sci* U S A 94, 3789-3794, doi:10.1073/pnas.94.8.3789 (1997).
15. DeKelver, R. C. et al. Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. *Genome Res* 20, 1133-1142, doi:10.1101/gr.106773.110 (2010).
16. Papapetrou, E. P. & Schambach, A. Gene Insertion Into Genomic Safe Harbors for Human Gene Therapy. Mol Ther 24, 678-684, doi:10.1038/mt.2016.38 (2016).
17. Rodriguez-Fornes, F. et al. Targeted gene therapy into a safe harbor site in human hematopoietic progenitor cells. *Gene Ther* 27, 435-450, doi:10.1038/s41434-020-0144-x (2020).
18. Pellenz, S. et al. New Human Chromosomal Sites with "Safe Harbor" Potential for Targeted Transgene Insertion. *Hum Gene Ther* 30, 814-828, doi:10.1089/hum.2018.169 (2019).
19. Arsenios Vlassis, T. L. J., Marina Mohr, Dominika J. Jedrzejczyk, Xiangyou Meng, Gergo Kovacs, Martı Morera-Gomez, Andrea Barghetti, Sergi Muyo Abad, Roland F. Baumgartner, Kedar N. Natarajan, Lars K. Nielsen, Tanya Warnecke, and Ryan T. Gill. CRISPR-Cas12a-integrated transgenes in genomic safe harbors retain high expression in human hematopoietic iPSC-derived lineages and primary cells. *iScience* 26, 108287, doi:https://doi.org/10.1016/j.isci.2023.108287 (2023).
20. Aznauryan, E. et al. Discovery and validation of human genomic safe harbor sites for gene and cell therapies. *Cell Rep Methods* 2, 100154, doi:10.1016/j.crmeth.2021.100154 (2022).
21. Odak, A. et al. Novel extragenic genomic safe harbors for precise therapeutic T-cell engineering. *Blood* 141, 2698-2712, doi:10.1182/blood.2022018924 (2023).
22. Shrestha, D. et al. Genomics and epigenetics guided identification of tissue-specific genomic safe harbors. *Genome Biol* 23, 199, doi:10.1186/s13059-022-02770-3 (2022).
23. Kimura, Y. et al. Human Genomic Safe Harbors and the Suicide Gene-Based Safeguard System for iPSC-Based Cell Therapy. *Stem Cells Transl Med* 8, 627-638, doi:10.1002/sctm.18-0039 (2019).
24. Kikuchi, Y. et al. Gene-modified blister fluid-derived mesenchymal stromal cells for treating recessive dystrophic epidermolysis bullosa. *J Invest Dermatol*, doi:10.1016/j.jid.2023.05.021 (2023).
25. Zou, J. et al. Oxidase-deficient neutrophils from X-linked chronic granulomatous disease iPS cells: functional correction by zinc finger nuclease-mediated safe harbor targeting. *Blood* 117, 5561-5572, doi:10.1182/blood-2010-12-328161 (2011).
26. Ordovas, L. et al. Efficient Recombinase-Mediated Cassette Exchange in hPSCs to Study the Hepatocyte Lineage Reveals AAVS1 Locus-Mediated Transgene

Inhibition. *Stem Cell Reports* 5, 918-931, doi:10.1016/j.stemcr.2015.09.004 (2015).
27. Klatt, D. et al. Differential Transgene Silencing of Myeloid-Specific Promoters in the AAVS1 Safe Harbor Locus of Induced Pluripotent Stem Cell-Derived Myeloid Cells. *Hum Gene Ther* 31, 199-210, doi:10.1089/hum.2019.194 (2020).
28. Urcelay, E., Ward, P., Wiener, S. M., Safer, B. & Kotin, R. M. Asymmetric replication in vitro from a human sequence element is dependent on adeno-associated virus Rep protein. *J Virol* 69, 2038-2046, doi:10.1128/JVI.69.4.2038-2046.1995 (1995).
29. Bertero, A. et al. Optimized inducible shRNA and CRISPR/Cas9 platforms for in vitro studies of human development using hPSCs. *Development* 143, 4405-4418, doi:10.1242/dev.138081 (2016).
30. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat Rev Genet* 16, 472-482, doi:10.1038/nrg3962 (2015).
31. Hildebrandt, E., Penzes, J. J., Gifford, R. J., Agbandje-Mckenna, M. & Kotin, R. M. Evolution of dependoparvoviruses across geological timescales-implications for design of AAV-based gene therapy vectors. *Virus Evol* 6, veaa043, doi:10.1093/ve/veaa043 (2020).
32. Penzes, J. J., Marsile-Medun, S., Agbandje-McKenna, M. & Gifford, R. J. Endogenous amdoparvovirus-related elements reveal insights into the biology and evolution of vertebrate parvoviruses. *Virus Evol* 4, vey026, doi:10.1093/ve/vey026 (2018).
33. Kapoor, A., Simmonds, P. & Lipkin, W. I. Discovery and characterization of mammalian endogenous parvoviruses. *J Virol* 84, 12628-12635, doi:10.1128/JVI.01732-10 (2010).
34. Callaway, H. M. et al. Examination and Reconstruction of Three Ancient Endogenous Parvovirus Capsid Protein Gene Remnants Found in Rodent Genomes. *J Virol* 93, doi:10.1128/JVI.01542-18 (2019).
35. Smith, R. H. et al. Germline viral "fossils" guide in silico reconstruction of a mid-Cenozoic era marsupial adeno-associated virus. *Sci Rep* 6, 28965, doi:10.1038/srep28965 (2016).
36. Feschotte, C. & Gilbert, C. Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* 13, 283-296, doi:10.1038/nrg3199 (2012).
37. Wang, Y. N. et al. The Role of Syncytin in Placental Angiogenesis and Fetal Growth. *Front Cell Dev Biol* 10, 852561, doi:10.3389/fcell.2022.852561 (2022).
38. Campbell, M. A., Loncar, S., Kotin, R. M. & Gifford, R. J. Comparative analysis reveals the long-term coevolutionary history of parvoviruses and vertebrates. PLoS Biol 20, e3001867, doi:10.1371/journal.pbio.3001867 (2022).
39. Martin, F. J. et al. Ensembl 2023. *Nucleic Acids Res 51*, D933-D941, doi:10.1093/nar/gkac958 (2023).
40. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat Genet* 48, 1193-1203, doi:10.1038/ng.3646 (2016).
41. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* 339, 823-826, doi:10.1126/science.1232033 (2013).

42. Oh, S. A. et al. High-efficiency nonviral CRISPR/Cas9-mediated gene editing of human T cells using plasmid donor DNA. *J Exp Med* 219, doi:10.1084/jem.20211530 (2022).
43. Bloomer, H., Smith, R. H., Hakami, W. & Larochelle, A. Genome editing in human hematopoietic stem and progenitor cells via CRISPR-Cas9-mediated homology-independent targeted integration. *Mol Ther* 29, 1611-1624, doi:10.1016/j.ymthe.2020.12.010 (2021).
44. Rai, R. et al. Targeted gene correction of human hematopoietic stem cells for the treatment of Wiskott - Aldrich Syndrome. *Nat Commun* 11, 4034, doi:10.1038/s41467-020-17626-2 (2020).
45. Cosenza, L. C. et al. Efficient CRISPR-Cas9-based genome editing of beta-globin gene on erythroid cells from homozygous beta(0)39-thalassemia patients. *Mol Ther Methods Clin Dev* 21, 507-523, doi:10.1016/j.omtm.2021.03.025 (2021).
46. Uchida, N. et al. Preclinical evaluation for engraftment of CD34(+) cells gene-edited at the sickle cell disease locus in xenograft mouse and non-human primate models. *Cell Rep Med* 2, 100247, doi:10.1016/j.xcrm.2021.100247 (2021).
47. Rai, R. et al. An improved medium formulation for efficient ex vivo gene editing, expansion and engraftment of hematopoietic stem and progenitor cells. *Mol Ther Methods Clin Dev* 29, 58-69, doi:10.1016/j.omtm.2023.02.014 (2023).
48. Six, E. et al. Clonal tracking in gene therapy patients reveals a diversity of human hematopoietic differentiation programs. *Blood* 135, 1219-1231, doi:10.1182/blood.2019002350 (2020).
49. Radtke, S. et al. Purification of Human CD34(+)CD90(+) HSCs Reduces Target Cell Population and Improves Lentiviral Transduction for Gene Therapy. *Mol Ther Methods Clin Dev* 18, 679-691, doi:10.1016/j.omtm.2020.07.010 (2020).
50. Seczynska, M., Bloor, S., Cuesta, S. M. & Lehner, P. J. Genome surveillance by HUSH-mediated silencing of intronless mobile elements. *Nature* 601, 440-445, doi:10.1038/s41586-021-04228-1 (2022).
51. Lee, B. C. et al. A RUNX1-FPDMM rhesus macaque model reproduces the human phenotype and predicts challenges to curative gene therapies. *Blood* 141, 231-237, doi:10.1182/blood.2022018193 (2023).
52. Cunin, P. N., P.A. Megakaryocytes as immune cells *Journal of Leukocyte Biology* 105, 1111–1121, doi:10.1002/JLB.MR0718-261RR (2019).
53. Nishikii, H., Kurita, N. & Chiba, S. The Road Map for Megakaryopoietic Lineage from Hematopoietic Stem/Progenitor Cells. *Stem Cells Transl Med* 6, 1661-1665, doi:10.1002/sctm.16-0490 (2017).
54. Zhang, J. et al. Non-viral, specifically targeted CAR-T cells achieve high safety and efficacy in B-NHL. *Nature* 609, 369-374, doi:10.1038/s41586-022-05140-y (2022).
55. Cabriolu, A. et al. Globin vector regulatory elements are active in early hematopoietic progenitor cells. *Mol Ther* 30, 2199-2209, doi:10.1016/j.ymthe.2022.02.028 (2022).
56. Liu, M. et al. Genomic discovery of potent chromatin insulators for human gene therapy. *Nat Biotechnol* 33, 198-203, doi:10.1038/nbt.3062 (2015).

57. Ogata, T., Kozuka, T. & Kanda, T. Identification of an insulator in AAVS1, a preferred region for integration of adeno-associated virus DNA. *J Virol* 77, 9000-9007, doi:10.1128/jvi.77.16.9000-9007.2003 (2003).
58. Shin, S. et al. Comprehensive Analysis of Genomic Safe Harbors as Target Sites for Stable Expression of the Heterologous Gene in HEK293 Cells. *ACS Synth Biol* 9, 1263-1269, doi:10.1021/acssynbio.0c00097 (2020).

Acknowledgements

This work was supported in part by funding from the Association Monégasque Contre les Myopathies (RMK), and the Bill & Melinda Gates Foundation (OPP1202116 to RMK).

Author contributions

R.M.K and M.Q-R designed experiments and analyzed data. M.Q-R. performed the experiments. S.L., M.A.C., and R.G analyzed genomes to map EPVs to the human genome. M.Q-R and M.A.C analyzed RNA-seq data. K.M.P. performed the CUT&RUN experiments. R.M.K. conceptualized and obtained funding. M.Q-R drafted the manuscript. All the authors revised and approved the final version.

Competing interests

This manuscript presents results that are covered by intellectual property co-owned by the University of Massachusetts Chan Medical School and Synteny Therapeutics. R.M.K. is a co-founder of Synteny Therapeutics, Inc.

Figure legends

Fig. 1. Identification of 102 potential genome safe harbor (GSH) sites in the human genome. **a.** Circle plot of EPV orthologues mapped onto the human genome. The outer layer represents human chromosomes, 1 to 22 and X. Chromosome lengths are indicated by the tick marks, e.g., 0 - 100 Mb. Black and red radial marks represent chromosomal banding patterns and centromeres, respectively. The inner layer depicts intergenic and intronic locations of GSHs as red and blue circles, respectively, with the estimated divergence time of the host species taxa from humans scaled from 0 to 350 million years ago (MYA). **b.** Comparative genomic analysis of the precision of homology between EPV host species and human genome. EPVs are derived from four parvovirus genera: *Amdoparvovirus*, *Dependoparvovirus*, *Erythroparvovirus* and *Protoparvovirus*. The distribution displays abundance of *Dependoparvovirus* and *Protoparvovirus* related loci. **c.** Propensity to genome editing of 17 selected human EPV orthologs evaluated by indel formation in HEK293T cells following the expression of *Cas9* and a specific gRNA. The intergenic or intronic category is depicted as red and blue bars, respectively. The nomenclature of the assessed GSH sites (abbreviated EPV name) used for this report as well as the chromosomal host are indicated on the x-axis. Indel frequencies were determined through TIDE decomposition analysis. Mean ± s.d., n = 2 independent experiments.

Fig. 2. Targeting of GSH candidates in primary human CD34$^+$ HSPCs. **a** Genome targeting with *eGFP* expression cassettes into GSH candidates using locus specific CRISPR/Cas9-ssDNA editing sets. Single-stranded DNA donor templates were flanked

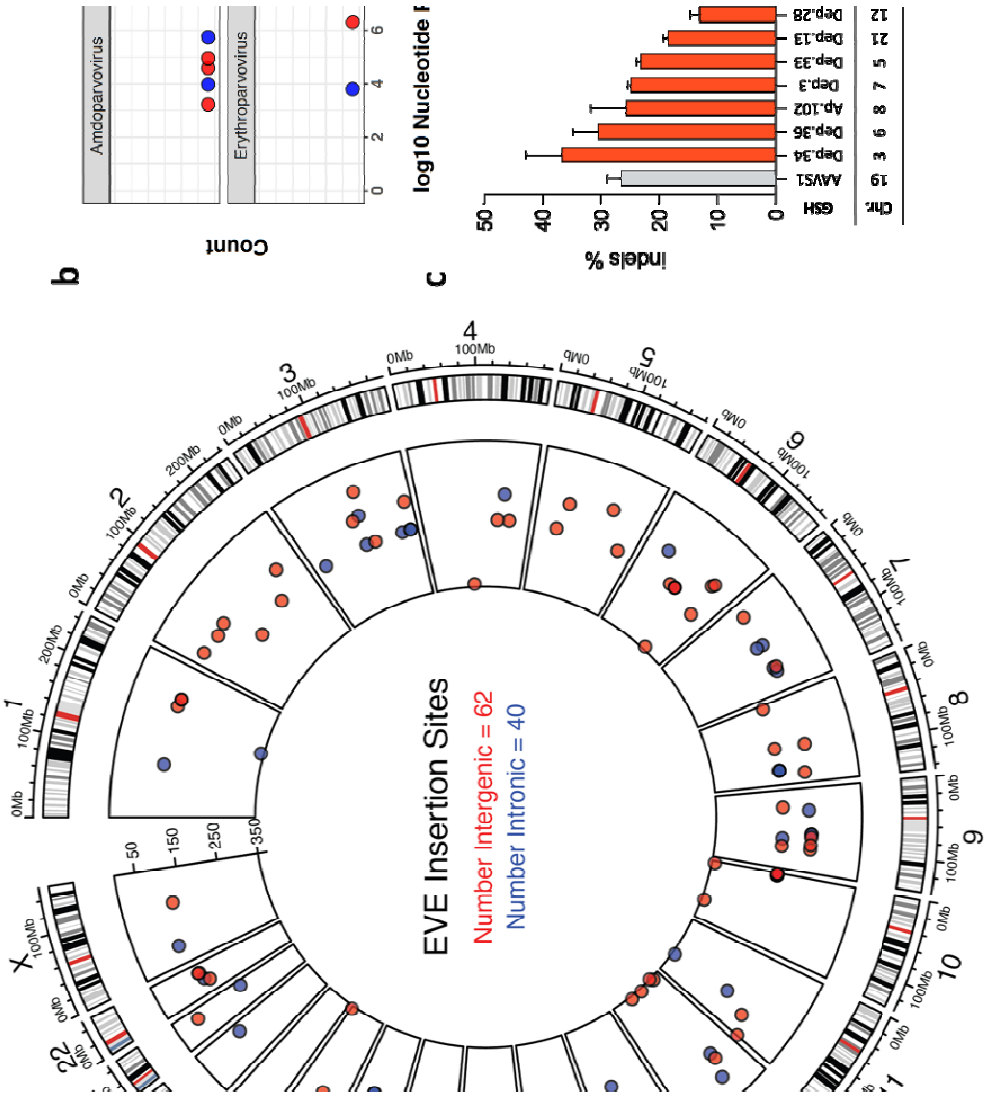
962 with 300 nt homology arms to mediate homologous recombination through single
963 stranded template repair (SSTR) pathway. **b** Experiments with bone marrow $CD34^+$
964 HSPCs. Pre-stimulated cells were nucleofected in presence of the CRISPR/Cas9-
965 ssDNA editing set and cultured for two days before the sorting of the GFP^+ population.
966 The GFP^+ cells were cultured for an additional three days before downstream
967 processing. Alternatively, AAV6 vectors were used to deliver templates to target *Ap.102*,
968 *Dep.33*, *Dep.13*, *Dep.1* and *Dep.35* loci. In parallel, 1 day-nucleofected cells were
969 cultured in erythroblast expansion conditions before maturation with human EPO. **c**
970 Percentage of GFP+ cells at two days post-nucleofection. Intergenic and intronic loci
971 are depicted as red and blue bars, respectively. Open circles represent individual
972 human donors. *AAVS1* GSH (gray) served as reference control. Mean ± s.d, ANOVA
973 followed by Dunnett test * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ and **** $P < 0.0001$. $n \geq$
974 4 independent experiments. **d** Cell viability determined by flow analysis of propidium
975 iodide stained cells. Comparison control corresponds to unedited cells (mock). Open
976 circles represent independent human donors. Mean ± s.d, ANOVA followed by Dunnette
977 test, $n \geq 3$ independent experiments. **e** Proliferation rate of GFP^+ HSPCs three days
978 after sorting. ANOVA followed by Dunnett test, $n \geq 3$ independent experiments. **f**
979 Targeting of long term-HSCs ($CD34^+/CD90^+$, magenta) and progenitor cells
980 ($CD34^+/CD38^+$, $CD34^+/CD38^-$) within the $GFP^+/CD34^+$ fraction. n = 2-3 independent
981 experiments using different human donors. **g** Targeted Integration in GFP+ cells sorted
982 after 14 days of liquid culture. Specific primer/probe sets (arrows and bars) were
983 utilized. The efficiency was calculated as the FAM/HEX ratio. $n = 2$ independent

experiments. **h** Persistent transgene expression after 45 days of liquid culture. **i** AAV6 delivering of templates for *Ap.102*, *Dep.33*, *Dep.13*, *Dep.1* or *Dep.35* loci show increased levels of sorted GFP^+ cells without affecting cell viability and high level of integration. Mean ± s.d., *n* = 2 independent experiments.

Fig. 3. **Gene expression changes following the targeting of GSH candidates in primary human HSPCs**. **a** Comparison of total differentially expressed genes (DEGs) among intergenic (*Dep.3*, *Ap.102*) and intronic (*Dep.2*, *Dep.55*, *Prot.218*) GSH candidates, red and blue symbols, respectively, compared to the reference GSH, *AAVS1,* shown in gray. Genes consistently upregulated (n = 30) and downregulated (n = 15) across all experiments are indicated by shading in barplots. **b** Volcano plots of the global transcriptional changes after genome editing of *AAVS1* (1008 up-/587 downregulated genes), *Ap.102* (133 up-/363 downregulated genes), *Dep.3* (220 up-/289 downregulated genes), *Dep.2* (1342 up-/821 downregulated genes), *Dep.55* (269 up-/656 downregulated genes) and *Prot.218* (798 up-/436 downregulated genes). Genes with adjusted *p*-values < 0.01 are highlighted in color, *AAVS1* gray, intergenic red and intronic blue. Highly significant DEGs are named. **c** Gene ontology annotations for DEGs obtained from *Dep.2*, *Dep.3*, *Ap.102*, and *Prot.218* editing. Genes with attributes to regulation of cell proliferation are indicated by yellow bars with representative genes plotted at the bottom, blue downregulated oncogenes, red upregulated tumor suppressors. **d** Heatmaps with the top 20 more significant DEGs per edited locus (all DEGs are available in a database uploaded on NCBI). The host gene for intronic loci or the flanking genes for intergenic loci are appended at the bottom. Color scale

represents the Log_2 of normalized counts of *n* = 2 independent donors (pink and green boxes).

Fig. 4. **Multipotency of modified CD34+ HSPCs is retained after gene addition**. **a** Colony forming units (CFUs) per targeted locus after 14 days in differentiation medium (bright field and epifluorescence images). The GSH locus designation is indicated on the top, intronic or intergenic category at the bottom, and CFU identity on the left (E, erythroid; G, granulocyte; GM, granulocyte–macrophage and M, macrophage). *AAVS1*, *Dep.2*, *Dep.3*, *Proto.218*, *Proto.2* and *Prot.181* GSHs allowed widespread transgene expression, whereas *Dep.55*, *Dep.1* and *Ap.102* restricted expression to CFU-G, -GM, and -M. Scale bar = 200 µm. **b** Total colony distributions per GSH-targeted cells. Mean ± s.d., *n* = 3 independent experiments. Two way ANOVA followed by Dunnett test * $P < 0.05$, ** $P < 0.01$. **c** Erythroid ($CD235a^+$) and immune ($CD69^+$) composition of GFP^+ expressing cells. Representative FACS plots are shown (left). Mean ± s.d., *n* = 2-3 independent experiments. **d** CUT&RUN QPCR assay with primers amplifying the MND promoter of the transgene shows presence of H3K4me3 activation mark on the transgene inserted in *Dep.55*, *Dep.33*, *Dep.1* or *Ap.102* GSH sites of K562-derived megakaryocytes (*n* = 3 experimental replicates). **f** Erythroid maturation from *AAVS1*, *Dep.2*, *Dep.3*, *Prot.181*, *Prot.2*, and *Prot.218* unsorted cultures. *Upper*, 14 days-erythroblasts retain GFP+ signal. Scale bar 100 µm. *Bottom*, Enucleation occurs in presence (+EPO) but not in absence of human EPO (-EPO). Nuclear counterstaining with Hoechst 33342. Representative pictures of *n* = 2 independent human donors.



1031
1032
1033

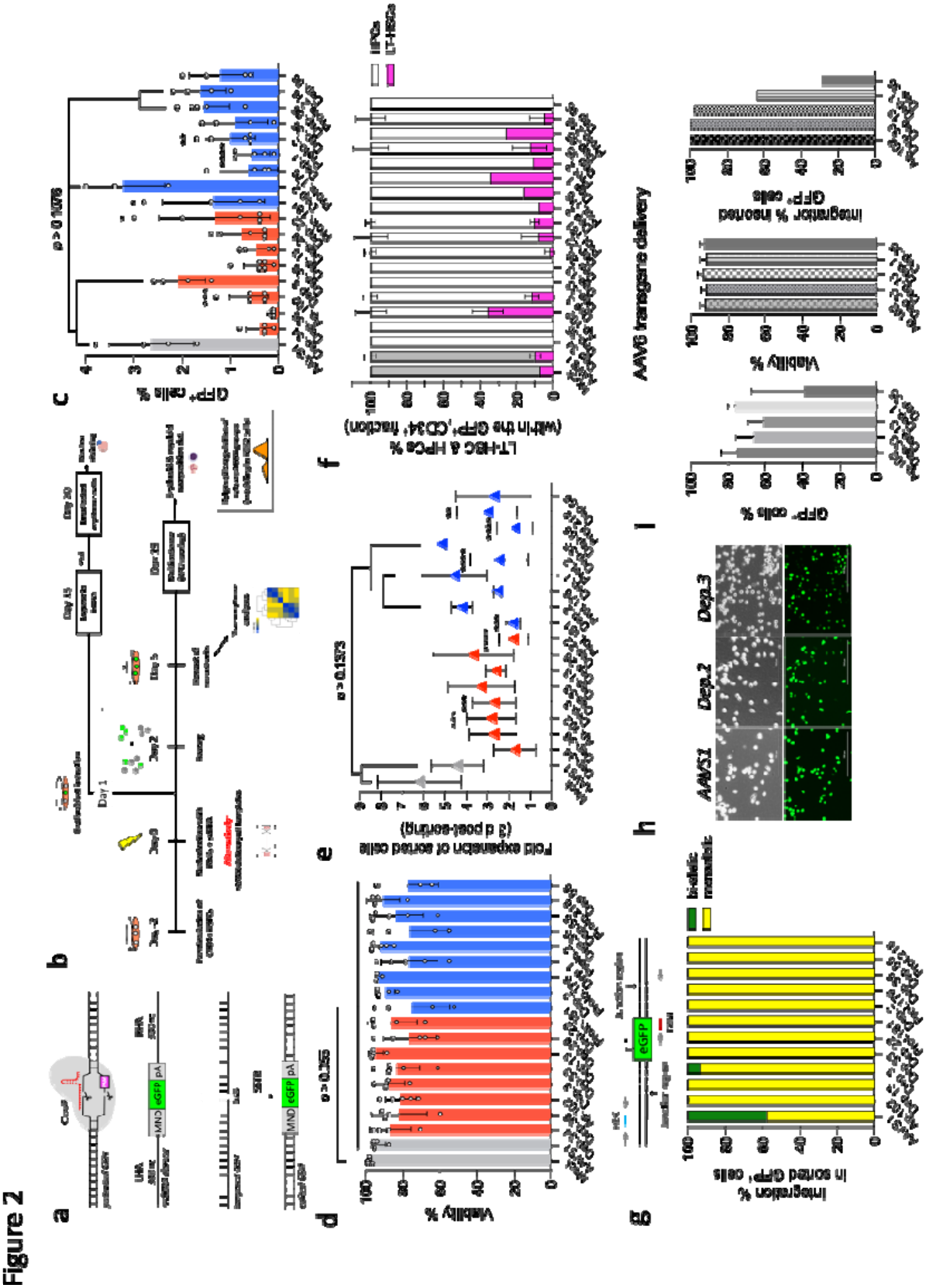
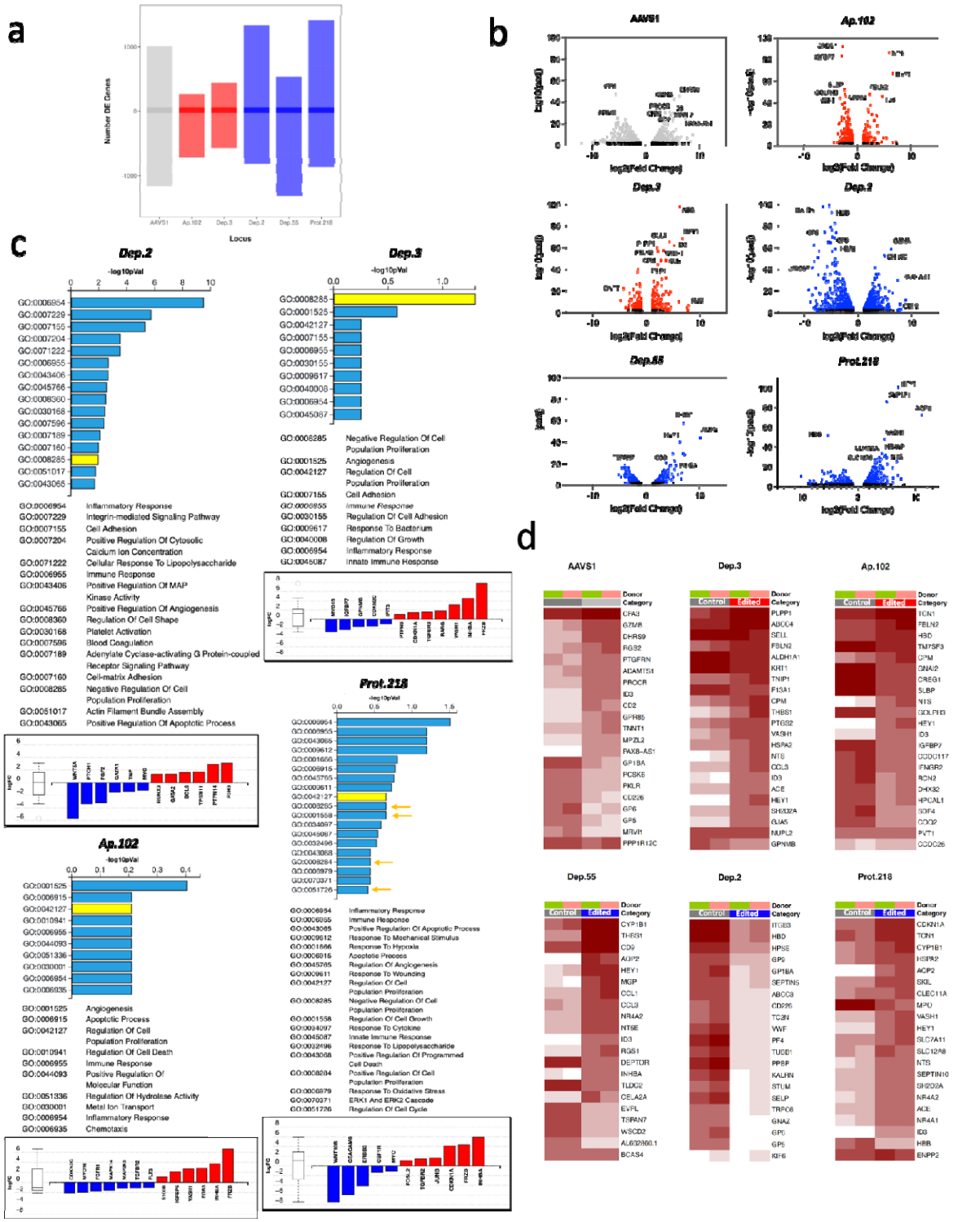
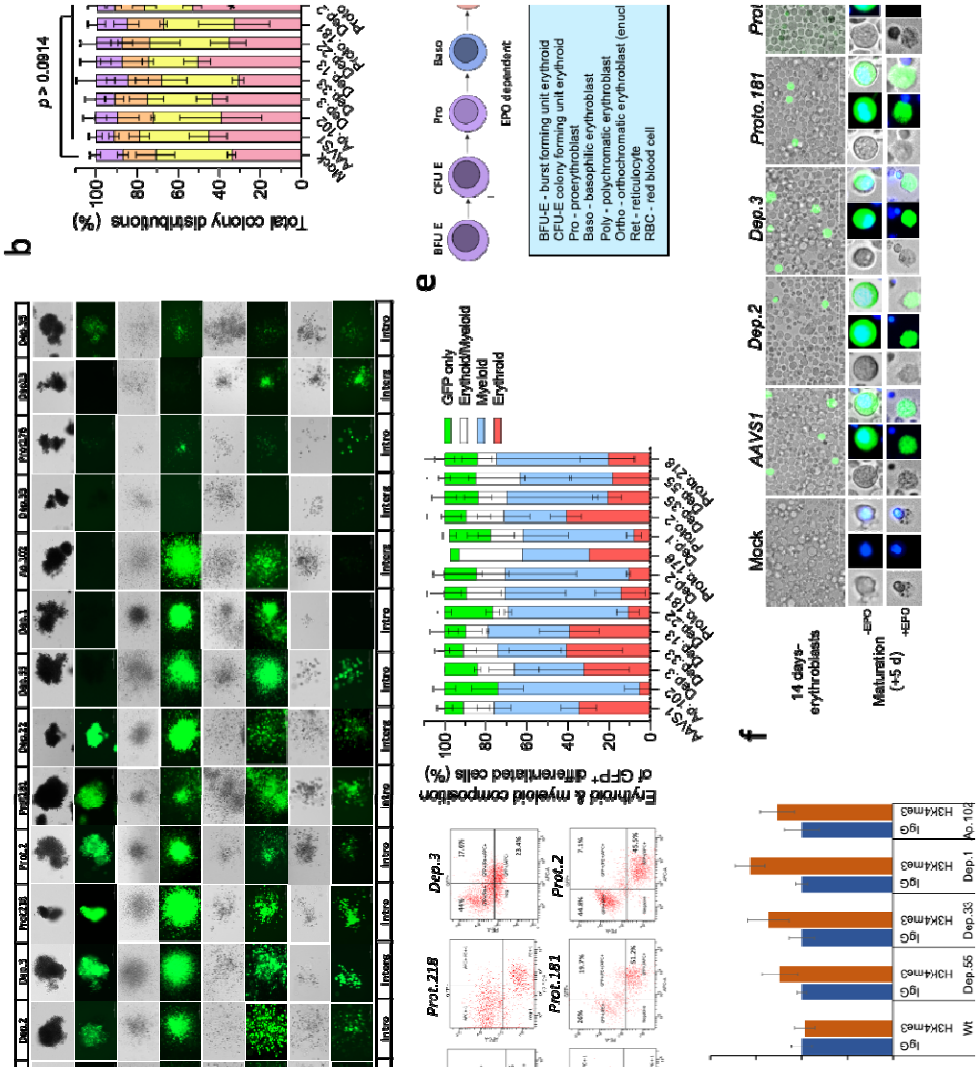


Figure 2
a
b
c
d
e
f
g
h
i
Day 1
Day 2
Day 3
Day 5
Day 15
Day 20
Day 28
GFP+ cells %
Viability %
Fold expansion of sorted cells (3 d post-sorting)
LT-HSC & HPCs % (within the GFP+,CD34+ fraction)
Integration % in sorted GFP+ cells
AAV6 transgene delivery
Integration % in sorted GFP+ cells
AAVS1
Dep.2
Dep.3
eGFP

Figure 3


a
Number DE Genes
Locus
AAVS1
Ap.102
Dep.3
Dep.2
Dep.55
Prot.218
b
c
Dep.2
-log10pVal
GO:0006954 Inflammatory Response
GO:0007229 Integrin-mediated Signaling Pathway
GO:0007155 Cell Adhesion
GO:0007204 Positive Regulation Of Cytosolic Calcium Ion Concentration
GO:0071222 Cellular Response To Lipopolysaccharide
GO:0006955 Immune Response
GO:0043406 Positive Regulation Of MAP Kinase Activity
GO:0045766 Positive Regulation Of Angiogenesis
GO:0008360 Regulation Of Cell Shape
GO:0030168 Platelet Activation
GO:0007596 Blood Coagulation
GO:0007189 Adenylate Cyclase-activating G Protein-coupled Receptor Signaling Pathway
GO:0007160 Cell-matrix Adhesion
GO:0008285 Negative Regulation Of Cell Population Proliferation
GO:0051017 Actin Filament Bundle Assembly
GO:0043065 Positive Regulation Of Apoptotic Process
Dep.3
-log10pVal
GO:0008285 Negative Regulation Of Cell Population Proliferation
GO:0001525 Angiogenesis
GO:0042127 Regulation Of Cell Population Proliferation
GO:0007155 Cell Adhesion
GO:0006955 Immune Response
GO:0030155 Regulation Of Cell Adhesion
GO:0009617 Response To Bacterium
GO:0040008 Regulation Of Growth
GO:0006954 Inflammatory Response
GO:0045087 Innate Immune Response
Ap.102
-log10pVal
GO:0001525 Angiogenesis
GO:0006915 Apoptotic Process
GO:0042127 Regulation Of Cell Population Proliferation
GO:0010941 Regulation Of Cell Death
GO:0006955 Immune Response
GO:0044093 Positive Regulation Of Molecular Function
GO:0051336 Regulation Of Hydrolase Activity
GO:0030001 Metal Ion Transport
GO:0006954 Inflammatory Response
GO:0006935 Chemotaxis
Prot.218
-log10pVal
GO:0006954 Inflammatory Response
GO:0006955 Immune Response
GO:0043065 Positive Regulation Of Apoptotic Process
GO:0009612 Response To Mechanical Stimulus
GO:0001666 Response To Hypoxia
GO:0006915 Apoptotic Process
GO:0045765 Regulation Of Angiogenesis
GO:0009611 Response To Wounding
GO:0042127 Regulation Of Cell Population Proliferation
GO:0008285 Negative Regulation Of Cell Population Proliferation
GO:0001558 Regulation Of Cell Growth
GO:0034097 Response To Cytokine
GO:0045087 Innate Immune Response
GO:0032496 Response To Lipopolysaccharide
GO:0043068 Positive Regulation Of Programmed Cell Death
GO:0008284 Positive Regulation Of Cell Population Proliferation
GO:0006979 Response To Oxidative Stress
GO:0070371 ERK1 And ERK2 Cascade
GO:0051726 Regulation Of Cell Cycle
d
AAVS1
Dep.3
Ap.102
Dep.55
Dep.2
Prot.218
Donor
Category
Control
Edited



Supplementary Table 2. Human GSH sites experimentally assessed in primary CD34+ HSPCs

EPV Nomenclature	Abbrev.	Category	Host gene	Human Coordinates	ATAC peak at
Dependo.34-megadermatvespertilion	Dep.34	Intergenic	*RPL18AP8/RCC2P5*	Chr3: 96,719,524-96,721,025	6.0 kb
Dependo.36-phyllostomidae	Dep.36	Intergenic	*TMEM200A/SMLR1*	Chr6:130,764,851-130,765,531	10.0 kb
Ap.102-procavia	Ap.102	Intergenic	*PVT1/CCDC26*	Chr8:128,435,359-128437048	29.69 kb
Dependo.3-lagomorpha	Dep.3	Intergenic	*NUP42/GPNMB*	Chr7:23,218,501-23,222,555	0.55 kb
Dependo.33-pteropodidae	Dep.33	Intergenic	*PARP8/ISL1*	Chr5:50,957,501-50,959,300	0.47 kb
Dependo.13-cercopithecidae	Dep.13	Intergenic	*HUNK/Mis18A*	Chr21:32,097,568-32,098,575	90.0 kb
Dependo.28-daubentonia	Dep.28	Intergenic	*LRRK2/CNTN1*	Chr12:40,658,392-40,660,499	31.64 kb

Dependo.22-laurasiatheria	Dep.22	Intergenic	*DIAPH2/PCDH19*	ChrX:98,175,245-98,175,812	16.0 kb
AAVS1 locus	AAVS1	intergenic	*PPP1R12C*	Chr.19:55,112,144-55,117,873	0.26 kb
Proto.181-PhaCin	Prot.181	Intronic	*MANCR*	Chr10:4,652,078-4,656,006	32.67 kb
Dependo.2-vespertilionidae	Dep.2	Intronic	*KIF6*	Chr6:39,422,982-39,425,107	15 kb
Proto.176-NanGal	Prot.176	Intronic	*POT1-AS1*	Chr7:125,056,326-125,057,364	5.14 kb
Dependo.1-whippomorpha	Dep.1	Intronic	*PAX5*	Chr9: 36,852,789-36,855,146	0.87 kb
Proto.2-MusSpr	Prot.2	Intronic	*PDZRN4*	Chr12:41,262,344-41,264,604	1.28 kb
Dependo.35-vespertilionidae	Dep.35	Intronic	*ANKRD7*	Chr7:118,262,500-118,263,554	24.6 kb
Dependo.55-rodent	Dep.55	Intronic	*BCAS4*	Chr20:50,799,100-50,801,360	4.0 kb
Proto.218-VomUrs	Prot.218	Intronic	*ENPP2*	Chr8:119,591,710-119,593,739	3.74 kb
Dependo.56-cavia	Dep.56	Intronic	*GUCA1C*	Chr3:108,912,857-108,913,858	1.14 kb

Footnote: The nomenclature established to identify human orthologs of EPVs presented in the first column refers to the parvoviral genus of origin. An abbreviated name for each locus is indicated in the second column. When a candidate GSH site falls in an intron, the host gene is indicated. For intergenic GSH sites the flanking genes are indicated. Human coordinates correspond to the sequences retrieved from Ensembl to engineer the gRNAs. The nearest ATAC peak (open chromatin spot) per candidate GSH site was determined with Integrative Genomics Viewer (IGV, software.broadinstitute.org/software/igv/) using the results informed by Corces M et al. *Nat. Genetics* 48 (2016) for CD34+ HSPCs. The GSH sites within a 2kb distance from a peak are highlighted in blue.

Supplementary Table 3. Guide RNAs selected per candidate GSH site

Target GSH site	CATEGORY	gRNA sequence (No PAM)
Dep.33	Intergenic	AACACCTAGCATTCACATAA
Dep.3	Intergenic	AGGCTTCGGCTTCCGCAACG
Dep.34	Intergenic	GTATGATGCAACTTGGTCCG
Dep.36	Intergenic	AGTAACCAAGGATTCCTGCT
Dep.22	Intergenic	CTTCTAACAAACCAAACATG
Ap.102	Intergenic	CCATTATTCATGGAGCACGC
Dep.28	Intergenic	TACTAATGCAGGTGTACATA
Dep.13	Intergenic	TCCCTGGTCCACCTTAAGGC
AAVS1	Intronic	GGGCCACTAGGGACAGGAT

Dep.56	Intronic	GCCCAGTGTAATCAAAGACA
Proto.2	Intronic	ATATAATTACTTAAGGCTGT
Dep.1	Intronic	GAGGCACCGTGATGCCGCAG
Proto.218	Intronic	TCCTCCGAACTCCACTAGAA
Dep.55	Intronic	GGGAAGATCACCTGGGCTCT
Proto.176	Intronic	GTACCTATCTGCCACAGTAC
Dep.2	Intronic	GTATACAATAATATGAAGAG
Dep.35	Intronic	GCACATACACTCATGGGTAG
Proto.181	Intronic	ACTATACTCCTTGCTCCCAC

Footnote: The GSH site *AAVS1* locates within the first intron of *PPP1R12C* gene on chromosome 19 and the corresponding *AAVS1* gRNA was as described (Mali P et al., *Science* 339 (2013)). Other gRNA sequences were engineered for this report based on gRNA prediction software. Three to five gRNA were experimentally evaluated per locus.

Supplementary Table 4. Candidate GSHs showing persistent transgene expression during differentiation experiments of blood cells

GSH	Category	GFP^+ %	P value	ATAC peak at
Dep.3	Intergenic	2.07 ± 0.54	0.8495	0.55 kb
Dep.22	Intergenic	1.32 ± 1.14	0.0258	16.0 kb
Ap.102	Intergenic	0.62 ± 0.41	< 0.0001	29.69 kb
Dep.2	Intronic	3.23 ± 0.86	0.8938	15.0 kb
Prot.181	Intronic	1.35% ± 1.05	0.0473	32.67 kb
Dep.1	Intronic	0.54 ± 0.36	< 0.0001	0.87 kb
Prot.2	Intronic	1.0% ± 0.51	0.0027	1.28 kb

Dep.55	Intronic	1.56 ± 0.53	0.1076	4.0 kb
Prot.218	Intronic	1.62 ± 0.53	0.1961	3.74 kb

Footnote: The percentage of GFP+ cells was determined by FACS sorting at 2 days post-nucleofection (mean ± sd). The nearest ATAC peak (open chromatin) per candidate GSH site was determined with the results informed by Corces M et al. *Nat. Genetics* 48 (2016) for CD34+ HSPCs as described in Supplementary Table 2. The GSH sites within a 2kb distance from a peak are highlighted in blue.

Supplementary Table 5. Reagents used during this research

Reagent	**Catalog number**
Anti-human CD34	BioLegend #343625
Anti-human CD69	BioLegend #310909
Anti-human CD235a	BioLegend #349105
Anti-human CD90 (Thy1)	BioLegend #328114
Anti-human CD38	BioLegend #303506
TruStain FcXTM	BioLegend #422302
IgG control antibody	EpiCypher #13-0042
Anti-H3K4me3	EpiCypher #13-0041

Anti H3k27me3	Activemotif #39055
SFEM II medium	SCT #09655
CD34+ Expansion Supp	SCT #02691
UM729	SCT #72332
Methocult	SCT# H4435
Erythroid Expansion Supp	SCT # 02692
Huma EPO	SCT # 78007
P3 primary solution	Lonza V4XP-3032
ArchiTect™ Cas9 nuclease	SCT #76004
ddPCR Supermix for probes (No dUTP)	BioRad #1863024
ssDNA Production System	Takara #632666

Supplementary Table 6. Primers used for ssDNA synthesis

Primer name	Sequence 5' > 3'	Modification
AAVS1 f	ATTCGGGTCACCTCTCAC	
AAVS1 r	ACGGCCGCGTCAGAGCA	5Phos
Dep.34	ATCACTGAAATCATAAATAGGATTTAGAAT	
Dep.34	TGCACTTGAACTGAAAGTTATTC	5Phos
Dep.36	ATTCAGTTGAACTCAGATAACTCGA	
Dep.36	TCAGTTTGTAAGTTTTAGATAGTTCC	5Phos
Ap.102 f	GCTCCTTTTCCCCAGTGG	
Ap.102 r	TGCTTCCAAGTTGCCTAAAAT	5Phos
Dep.3 f	ACGCTAATGTCTTACTGAAATATAC	
Dep.3 r	CTGGGTCTGGGAACACCAC	5Phos

Dep.33 f	TGAGTTTGATGGAGAAAGATGGT	
Dep.33 r	CTGCAGTCCACTAATGAATTAATAG	5Phos
Dep.13 f	ACAGTTCCACGTGGCTGGG	
Dep.13 r	CAGGAAGCATCATTGTCTCC	5Phos
Dep.28 f	GGTAAGCCAGAGCTTTTTCTACT	
Dep.28 r	TACACTGTTGTTGTTGTTGTTTTTTTT	5Phos
Dep.22 f	CCAGCCAGCCTTAAAATTAGAT	
Dep.22 r	TGCAACTAGGTATTTAAACTCAATTCC	5Phos
Prot.181 f	AAATTTATTTTTTATCTGTTGTGAACTAGAGATC	
Prot.181 r	TTCCAGCCTCTAGAACTATGAG	5Phos
Dep.2 f	TCCTGATGGTGCCACACGG	
Dep.2 r	CAATGCACTCAGCCCCATG	5Phos
Prot.176 f	AAACAAATTTTGTAATTCATGCATC	
Prot.176 r	TCCTCATATGAAAATTTCTGCAAAT	5Phos
Dep.1 f	AGAAAAACCGAGAGAGAGGG	
Dep.1 r	GATTAAACTTGGACCAGTGTTGTTC	5Phos
Prot.2 f	ACCACTACATGATAAAAAGAAAATAAGA	
Prot.2 r	CCACACCTTAAAATACAAGAAGGAG	5Phos
Dep.35 f	GACAGGAAATTTATTAGACAGTTTACAGC	
Dep.35 r	TTCTGGTGCACTAAGATTTTAACT	5Phos
Dep.55 f	TTCTCCTCAGCCAGCCAA	
Dep.55 r	GCCTCGGTATGTGCCAAG	5Phos
Prot.218 f	GGGCAACTTAAATTGAATCAGCT	
Prot.218 r	CCCCGAATCCACTGAGGGAA	5Phos
Dep.56 f	GTCAGTGTACAGTCAACAGCA	
Dep.56 r	CTCAGCCTCCCAAAGTGC	5Phos

Footnote: Primers were used to amplify the 2 kb cassettes from the corresponding plasmid. The reverse primers were 5'-phosphorylated (5Phos) to facilitate the targeted degradation of the reverse strand by sequential strandase-treatments yielding ssDNA templates.

Supplementary Table 7. Primers to determine GSH-specific integration through ddPCR

Primer name	**Sequence 5' > 3'**	**Amplicon size (bp)**
Universal	GCCTGAGGATCCGATCTTTT	common fwd primer
AAVS1	TCACAGGTAAAACTGACGCA	595
Dep.3	TGAATACCTCTGCAAAGGGC	614
Dep.2	AGGTGGTCTAGAACTCCTGG	533
Dep.34	TGCTCAGTACCTAAGCTTTGG	500

Dep.36	CATTGGCCTCAGTCAGTACC	477
Ap.102	AGTCTTTCTCACTGCTGACA	656
Proto.181	CAAGAAGGCCCTCACAGATG	487
Proto.176	CCTCTGTGTCACTTTCACTGT	534
Dep.35	CGGGTAAATGTTAGGTGCCT	638
Dep.33	ACCATGTGTAACTGCAGTCC	463
Dep.1	GGGTCATTTGTACATGCCCA	535
Proto.1	AGATCCCACCACTGCACT	656
Dep.13	TTCATGCTTGTCAGAGACCG	476
Dep.22	ATGTGAACTGCTGGCCTTTT	475
Dep.28	GAGGTCTTTCTATGGTGCCA	551
Dep.56	ACGGGGTTTCACCATATTGG	513
Proto.218	AAAACGAGCCCATGGAGTTT	560
Dep.55	TGGAGCAGCAGAAATCTCAG	543
Sequencing	CTGCCAAAAATTATGGGGACA	NA

Footnote: The Universal primer is complementary tothe transgene and was includedas a forward primer for PCR reactions. This is paired with a locus specific reverse primer which aligns outside of the recombination boundary of the targeted locus. The PCR product results from locus specific integration. Sequencing primer aligns the rabbit poly(A) signal and was used to confirm the targeted integration of the transgene through Sanger-sequencing of purified amplicons.

Supplementary Table 8. Reference primers recognizing an unedited region of the candidate GSHs

Primer pair name	Sequence 5' > 3'	Targeted sequence	Amplicon (bp)
AAVS1 f	CTCTGACCTGCATTCTCTCC	chr19:55116021-55116564	544
AAVS1 r	TGCCCAAATGAAAGGAGTGA		
Dep.3 f	CAAGCATACATATGGCACGA	chr7:23220274-23220830	557
Dep.3 r	GACCCAGGTCTCCTTTTCTA		
Dep.34 f	AAAGCCGAGACAGGTTGAAA	chr3:96718554-96719078	525
Dep.34 r	GTTTGATGCCTTGAAGCCAG		

Dep.36 f	TTAATGCTTCTGGGAAAGAATTT	chr6:130764868-130765268	401
Dep.36 r	GCCTTGCTTCAATTCCATGA		
Ap.102 f	CTGAAGCAGGAACAGGAGAG	chr8:128435204-128435765	562
Ap.102 r	GCGGGTGTGAGATTTACCAT		
Dep.2 f	GACTCAGGTATCCAACAGCC	chr6:39423568-39424112	545
Dep.2 r	TGGGGGATGGAAGAGAAGAA		
Prot.181 f	TACGAGATAAGGGGCTCCTC	chr10:4652816-4653337	522
Prot.181 r	AACAGGGAGGCTTTTACGAC		
Prot.176 f	CCAGGAACTCTGCATGACAT	chr7:125055768-125056271	504
Prot.176 r	CTTGCCTCTCTCAGAAGCAG		
Dep.35 f	TCCCGAGTAGCTGAGACTAC	chr7:118262291-118262865	575
Dep.35 r	GAGGCATGATTAAGGCACCT		
Dep.33 f	ACTTCAAAGCCATGAACCAT	chr5:50957660-50958167	508
Dep.33 r	GCTTTCTTCTTAAGCTCACCA		
Dep.1 f	CCGATATACACGTACACATGC	chr9:36854330- 36854894	565
Dep.1 r	CGGGAGAAACAGAAGTGAGA		
Prot.2 f	TGGTGTTGCATGCCTGTAAT	chr12:41263414- 41263975	562
Prot.2 r	TCATGGAGTCAGTTTTGGGC		
Dep.13 f	GTCTTTTCTCTCTTCCACAAGA	chr21:32097016-32097537	522
Dep.13 r	ACTGTGAGTCCATTAAACCTC		
Dep.22 f	GCTCCTTCTCTACCACTTTATT	chrx:98174513-98175022	510
Dep.22 r	GAAATGTGACTATTTCATCTATACATT		
Dep.28 f	AGATGCGGGTTTTACCATGT	chr12:40658161- 40658733	573
Dep.28 r	TTGAGATTGCTGCAAATGCC		
Dep.56 f	CTGTTCACCAGCACATTTGT	chr3:108913558-108914068	511
Dep.56 r	TTTAAAATGGAGTCTTGCTTTGT		
Prot.218 f	TGGTGAAAGCCCATGTCTAC	chr8:119592290-119592807	518
Prot.218 r	CTCCTCTTGAGTGGGCTAGA		
Dep.55 f	AGCATCCTGAGTGCATTTCT	chr20:50799100- 50799609	510
Dep.55 r	GGTTGGCTGGACTGAGATTT		

Supplementary Table 9. Fluorescent probes for ddPCR

Probe name	Sequence 5' > 3'	Fluorophore
Universal	TGAAGCCCCTTGAGCATCTGACTTCTGGCT	FAM
AAVS1	CCCTTGCGTCCCGCCTCCCCT	HEX
Dep.3	ACAGCCAGAGGCAAAAGCTTGCACGC	HEX
Dep.34	CAGCCACAACATCCCCATAAGCCAAAGCCT	HEX
Dep.36	TGTGCAGGGACATAGACCCTCCTCCTCCT	HEX
Ap.102	ACAACCCACCACCGCTATGCAGCAAGA	HEX

Dep.2	CGCACTGCCTGTCCGCAGGCT	HEX
Prot.181	TGCGAAGCTGCCTGGGGCGC	HEX
Prot.176	AGTCCCACGGCATGGATCCTGTGGATAACA	HEX
Dep.35	CCGCCTGCCTCAGCCTCCCA	HEX
Dep.1	GCTCAGGCCAGCCACGAGGGC	HEX
Dep.33	ACCTTCCCGCTGCTGCTGCTGAGT	HEX
Prot.2	ACCACTGCACTCCAGCCTGGGCA	HEX
Dep.13	CCCACCACTCCCCGCCCCCT	HEX
Dep.22	TGCAGCATCTGGCATACTTCCTGGCC	HEX
Prot.181	GCCTGGGGCGCTAGCAGGCA	HEX
Dep.28	GCCACCATGCCTGGCCTCAGT	HEX
Dep.56	TGGGGCTGGGCACGGTGGCT	HEX
Prot.218	ACCCGGGAGGTGGAGGTTGCA	HEX
Dep.55	ACCAGCCTGTCTCCCGGGGGC	HEX

Footnote: The Universal probe (FAM), complimentary to the transgene, was used to quantify modified, i.e., edited, alleles. The HEX probes are specific for the unedited region of the targeted locus. The combined results were used to determine allelic insertions in the GFP^+ cells.

Supplementary Table 10. Primers for CUT&RUN qPCR

Primer pair name	Sequence 5’ > 3’	Amplicon (bp)
MND promoter f	CCATCCACGCTGTTTTGACC	71
MND promoter r	GCCCTTGCTCACCATGACTA	
ACTB f	CGGGGTCTTTGTCTGAGC	128
ACTB r	CAGTTAGCGCCCAAAGGAC	
HOXD8 f	GGCGAGGCCATCAATCCCAC	101
HOXD 8 r	CTGTTGCCATAGAGCTGCAGG	

Supplementary Figure 1

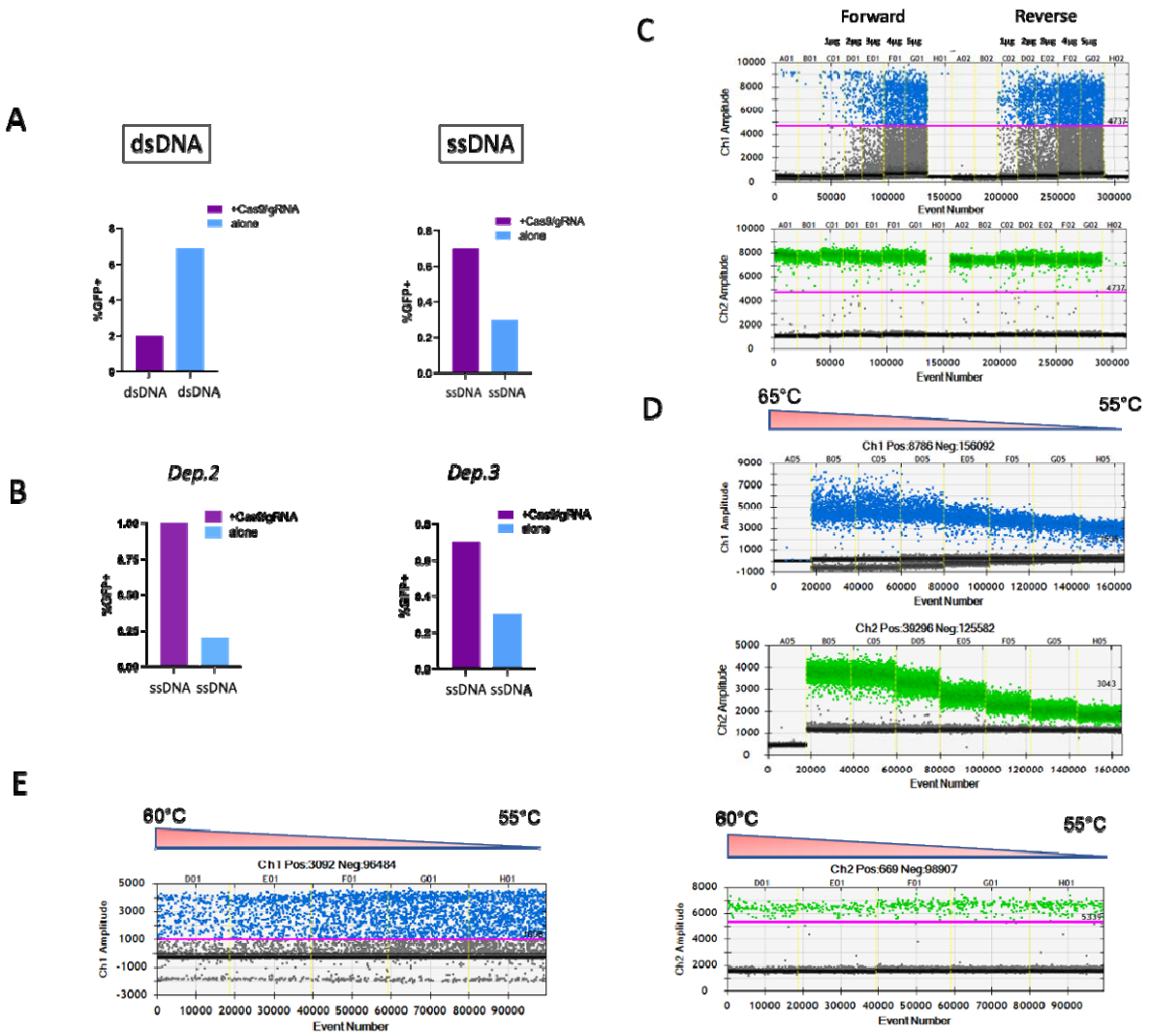


Fig. S1. Comparison of dsDNA and ssDNA donor templates for genome editing of human CD34+ HSPCs. A) Two days pre-stimulated CD34$^+$ HSPCs were transfected with dsDNA (HITI design) or ssDNA donor templates (HR), both carrying the same *eGFP* cassette, in presence or absence of CRISPR/Cas9 complexes specific for *Dep.2* locus. Two days after the nucleofection the GFP$^+$ cells were sorted by FACS. Nucleofection with dsDNA alone (i.e., in absence of CRISPR/Cas9 complexes) resulted in a greater number of GFP$^+$ cells than cells including CRISPR/Cas9 complexes. In contrast, cells treated with ssDNA alone showed less percentage of GFP$^+$ cells in

absence of CRISPR/Cas9 complexes indicating that ssDNA itself prevents GFP expression from non-integrated templates and suggests that the likelihood of random integration of ssDNA occurs infrequently, perhaps due to reduced stability of the ssDNA template. B) The same phenomenon was observed independently of the targeted locus (*Dep.2* and *Dep.3* are shown). C) Nucleofection of dsDNA templates also affect the quantification of edited cells through ddPCR as shown by 1D plots. The GFP^+ cell populations, processed by ddPCR, displayed the same rain pattern with increasing amounts of dsDNA (1-5 µg). The primer sets to detect the transgene inserted in forward or reverse orientation (following the HITI approach) produce a similar rain effect (blue droplets) in Quantasoft. Conversely, primers recognizing an unedited region (green droplets) produce defined droplet clusters. D) 1D plots from $CD34^+$ HSPCs transfected with ssDNA templates targeting *Dep.2* locus. Primers amplifying the junction regions (as described in main text) and a reference region generate defined clusters for quantitation even at different temperatures. E) Delivery of templates through rAAV6 produces as well an increased noise/background (rain effect) as shown by targeting of the *Dep.13* locus. Notice that the reference amplimer (right) did not show changes under the same temperature gradient.

Supplementary Figure 2

Fig. S2. GSH-specific integration using ssDNA templates. Amplicons obtained from edited primary HSPCs (oligos available in Supplementary Table 6) were gel purified and Sanger-sequenced with a common primer annealing the rabbit poly(A) signal. All amplimers started with the poly(A) sequence included in the transgene (green) followed by the right homology arm used for targeted recombination (magenta). The arrow indicates the junction between the right homology arm and the corresponding targeted locus (black) showing continuity among the sequences.

Supplementary Figure 3

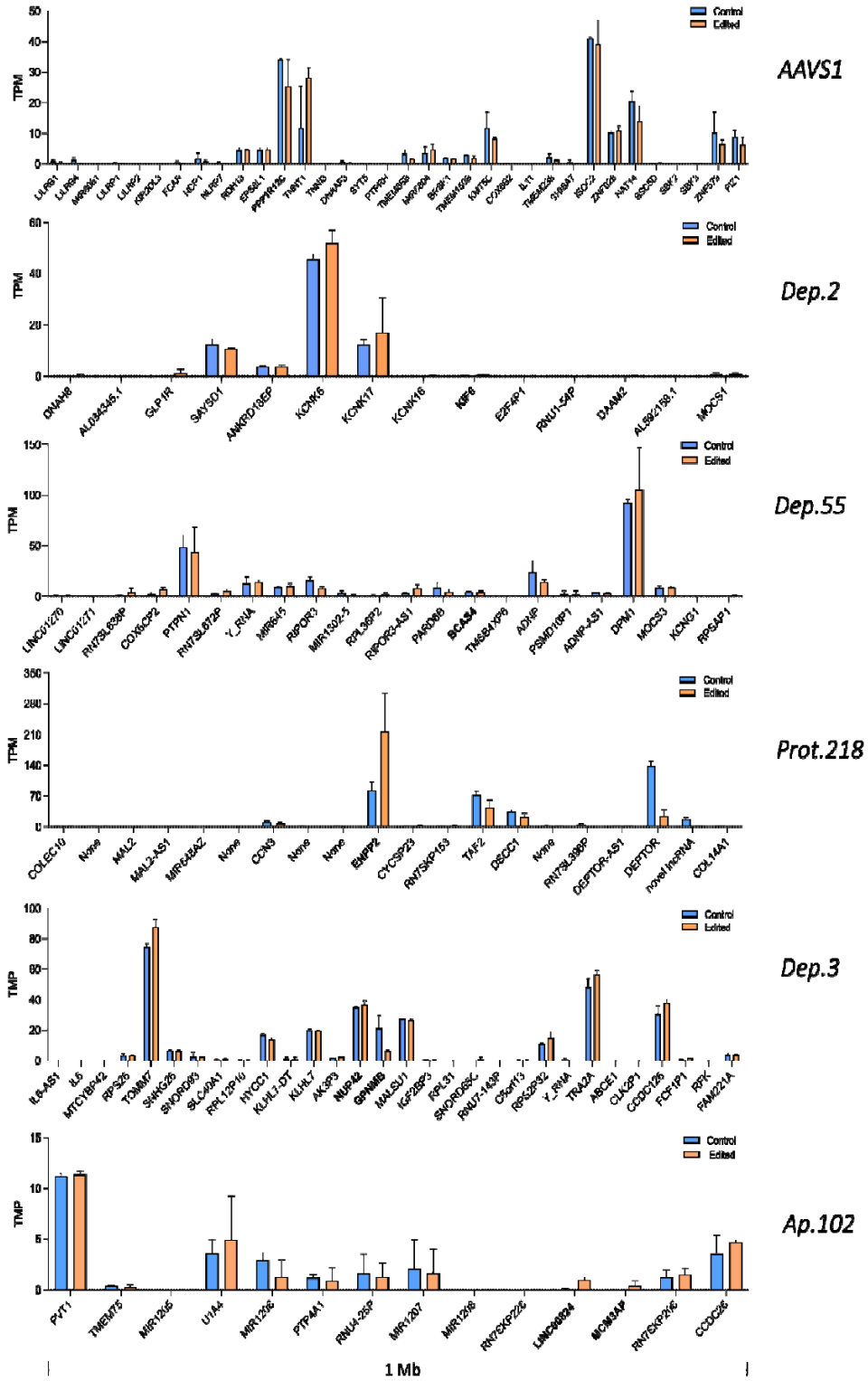


Fig. S3. **Gene expression changes within 1 Mb surrounding the insertion sites.** We used normalized expression values (transcript per million, TPM) from RNAseq to depict transcriptional changes within 1 Mb around the insertion sites. For intronic GSH sites the host gene is bold. For intergenic GSH sites the flanking genes are bold. The comparison before (blue) and after (orange) gene addition displayed minimal or no disruptions of transcriptional units surrounding *AAVS1*, *Dep.2*, *Dep.55*, *Prot.218*, *Dep.3* and *Ap108* GSH sites. In the *Proto.218* locus a fair up-regulation of the host gene (*ENPP2*) was detected. Mean ± s.d. n = 2 independent human donors.

Supplementary Figure 4

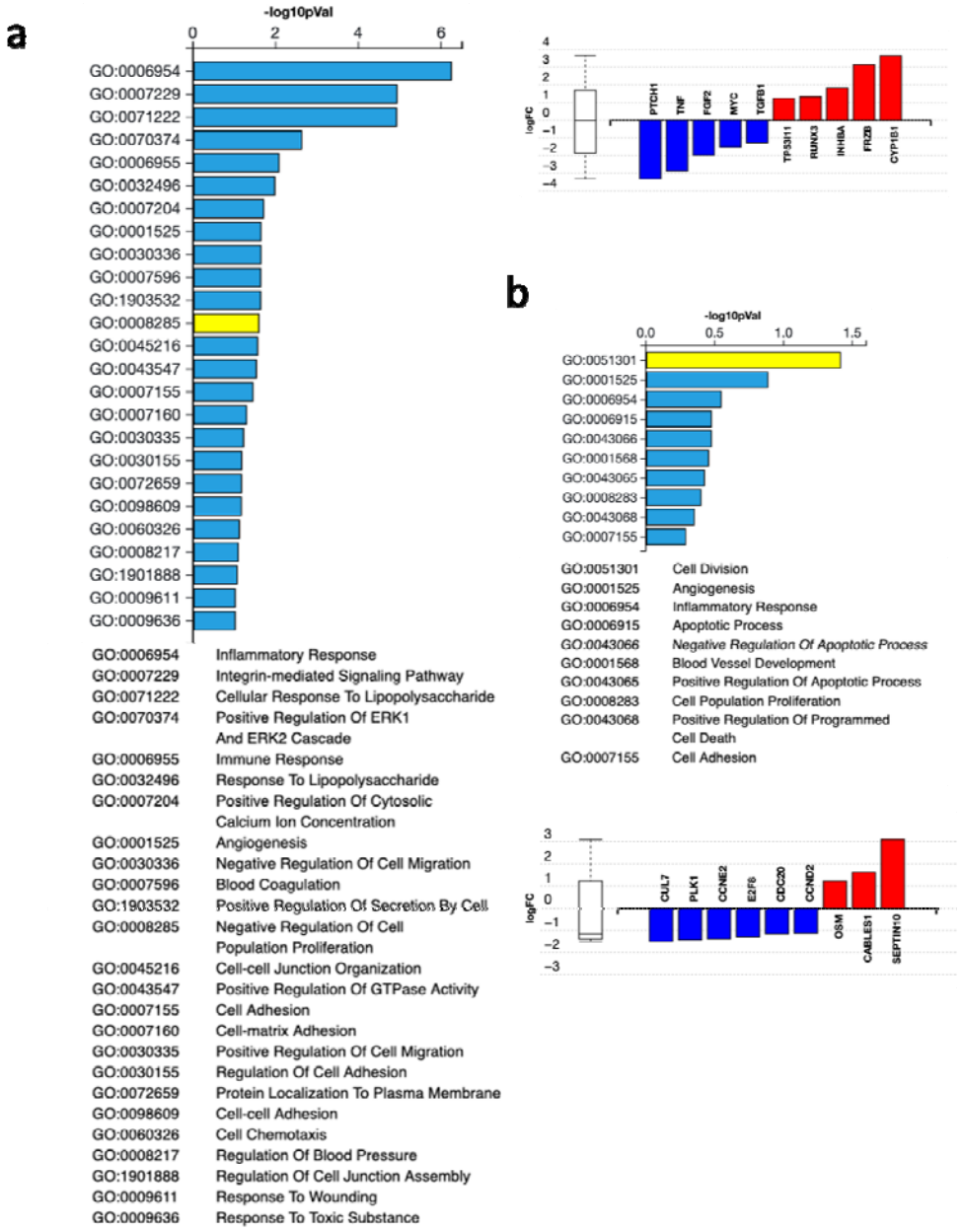


Fig. S4. **Gene ontology annotations for *AAVS1*- and *Dep.55*-targeted cells. A** *AAVS1*. **b** *Dep.55*. Both GSH-edited cells include annotated genes to regulate cell proliferation (yellow), however, for Dep.55 GSH site those genes represent the first featured group suggesting tight regulation of cell proliferation in vitro.

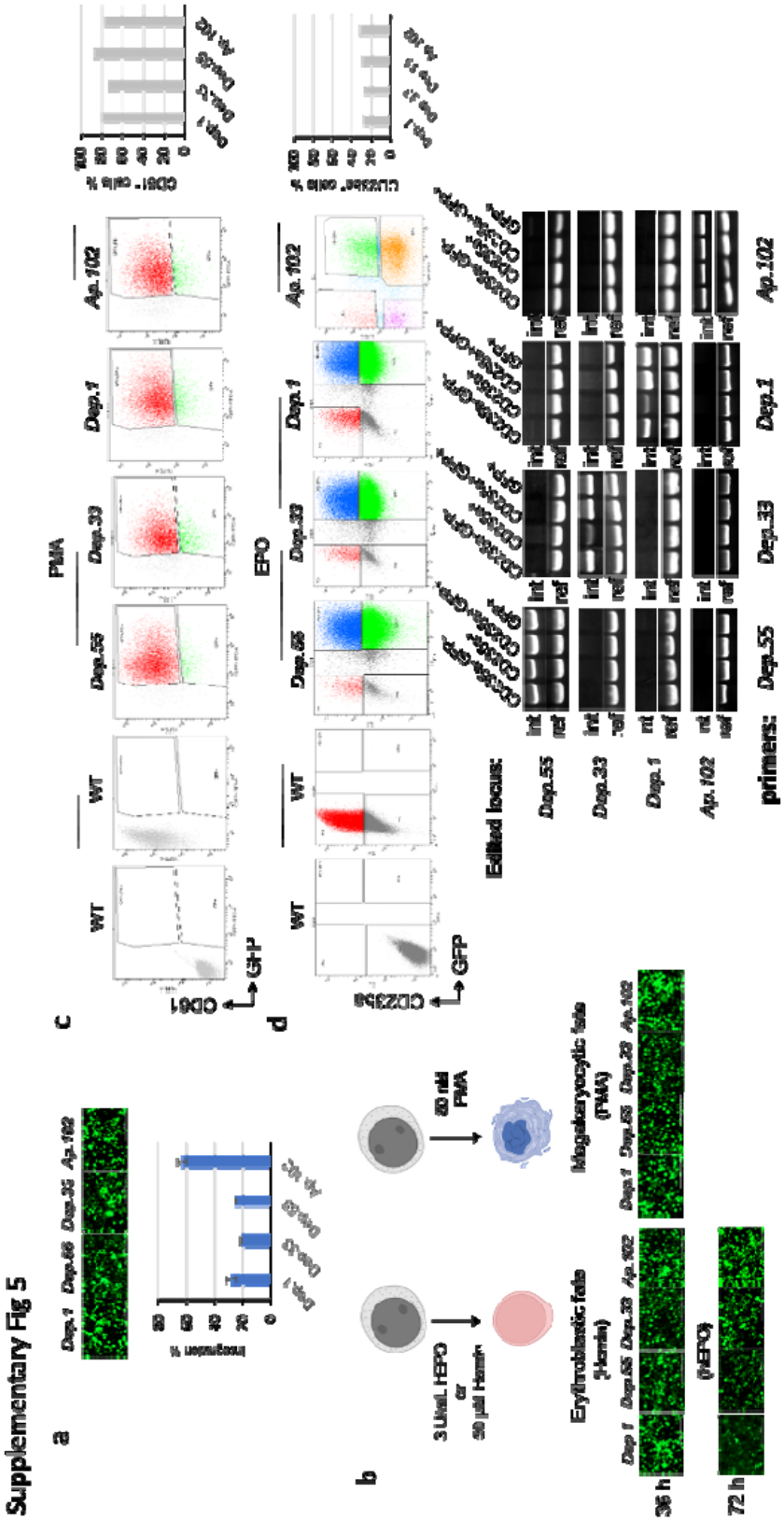

Supplementary Fig 5
a
b
c
d
PMA
EPO
WT
Dep.55
Dep.33
Dep.1
Ap.102
GFP
CD61
CD235a
Edited locus:
primers:
Erythroblastic fate (Hemin)
Megakaryocytic fate (PMA)
36 h
72 h
(hEPO)

Fig. S5. Erythroid and megakaryocytic induction of *Dep.1*, *Dep.55*, *Ap.102*, or *Dep.33* knocked in K562 cells. **a.** Stable cells were generated through delivery of ssDNA templates and the appropriated CRIPSR:gRNA complexes. The percentage of edited alleles was determined through ddPCR similarly to $CD34^+$ HSPCs (Supplementary Table 10). Given that K562 is triploid, a 30% of integration indicates mostly monoallelic integration. **b** Modified K562 cells were induced to erythroid or megakaryocytic phenotypes by stimulus with 3 U/mL hEPO (or 50 µM Hemin) and 80 nM PMA, respectively. Representative pictures of induced cells at indicated times are showed. **c** Cell sorting detected up to 90% of CD61+ megakaryocytic cells with sustained expression of *eGFP* (representative dot plots and corresponding bar plot are showed). **d** Erythroid induction reached 30% $CD235a^+$ cells after 3 days of hEPO treatment. Sorted phenotypes (i.e., $CD235a^-GFP^-$, $CD235a^+$, $CD235a^+GFP^+$, and GFP^+) were processed through PCR to confirm the presence of the transgene and discard effects produced by potentially random integrated clones (below). As shown the transgene was detected according to the stably edited K562 cell line (i.e. *Dep.1*, *Dep.33*, *Dep.55* or *Ap.102*), whereas the reference amplicon was observed independently of the analyzed sample. “int” refers to integration primers, and “ref” corresponds to reference primers which sequences are available in Supplementary Table 7).

Supplementary Figure 6

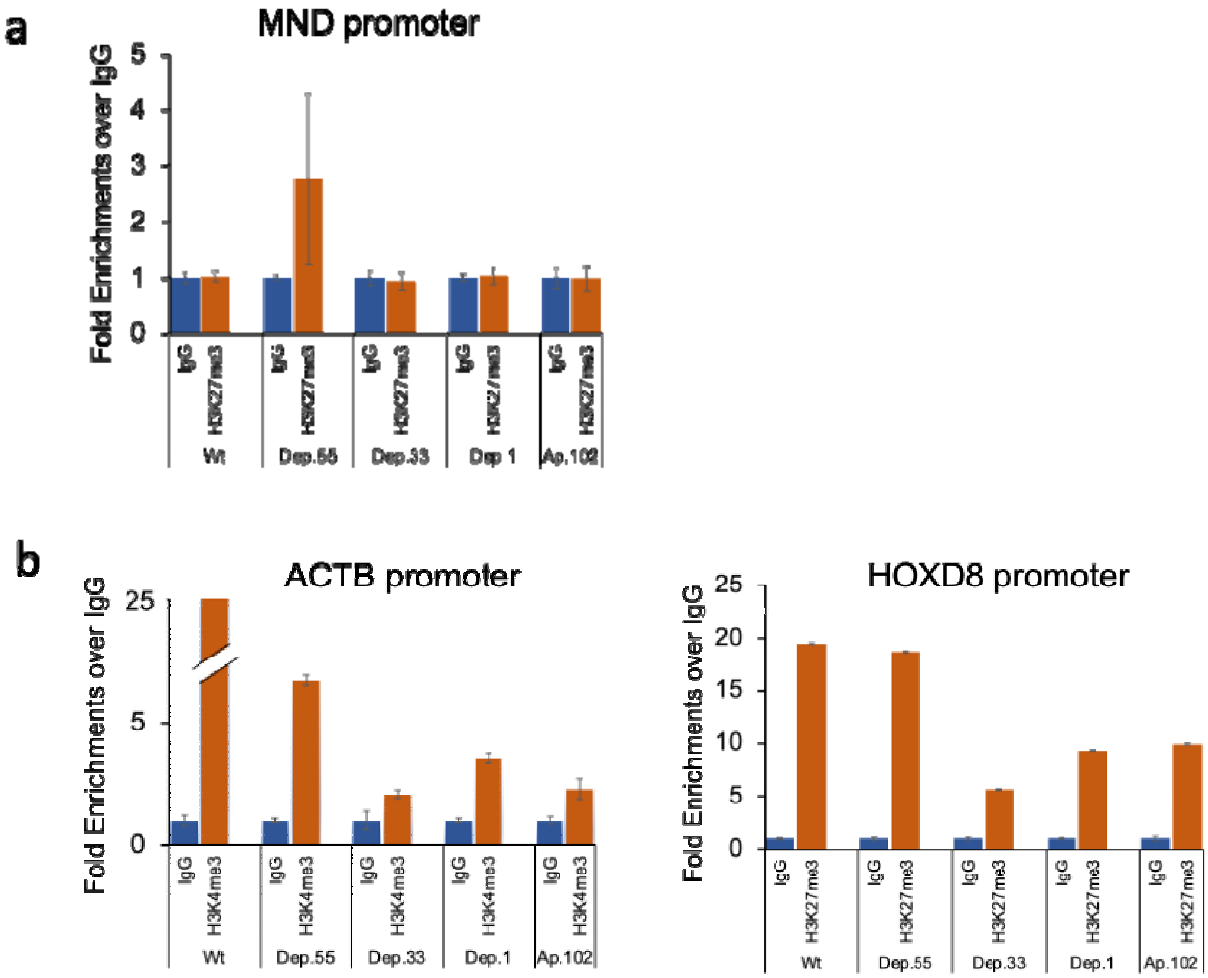


Fig. S6. H3K27me3 shows no enrichment at the transgene in induced megakaryocytic cells a.) Enrichment of H3K27me3 at the MND transgene promoter. b) Positive control enrichments of H3K4me3 at the *ACTB* locus and H3K27me3 at the HoxD8 locus