Identification of genome safe harbor loci for human gene therapy based on evolutionary biology and comparative genomics

Marco A. Quezada-Ramírez[1,2], Shannon Loncar[1,2,3], Matthew A. Campbell[4,5], Robert J. Gifford[6], and Robert M. Kotin[1,2,3,*]

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

[2] Microbiology and Physiological Systems
University of Massachusetts, Medical School

[3] Current affiliation Carbon Biosciences,
880 Winter Street, Waltham, MA 02451, USA

[4] University of Alaska Fairbanks
1731 South Chandalar Dr. Fairbanks, AK 99775, USA

[5] University of California, Davis
1 Shields Ave, Davis, CA 95616, USA

[6] University of Glasgow
Glasgow G12 8QQ, UK

*Correspondence to Robert.Kotin@umassmed.edu

Robert M. Kotin 0000-0001-5408-2476
Matthew A. Campbell 0000-0002-5826-0329
Marco A. Quezada-Ramírez 0000-0001-7490-4959
Robert J. Gifford 0000-0003-4028-9884

Abstract

Gene transfer into hematopoietic stem and progenitor cells (HSPCs) involving indiscriminately integrating viral vectors has unpredictable outcomes including potential adverse events such as leukemogenesis, resulting from insertional mutagenesis. Therefore, identifying and characterizing genomic safe harbor (GSH) sites where exogenous genetic information can be integrated safely 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 element (EPV) integration events identified in host genomes, 102 loci were mapped to the human genome with 17 being evaluated for potential use as GSHs in primary human CD34+ HSPCs. Six of the edited loci resulted in sustained transgene expression in both erythroid and myeloid phenotypes while three exhibited myeloid specific-expression and the remaining four loci resulted in a subset of myeloid and erythroid lineages. Following this approach, a minimum of nine promising GSH sites suitable for gene therapy of blood disorders were identified and additional GSH sites are likely to emerge from the remaining mapped loci which may be suitable for gene addition in stem and progenitor cells other than hematopoietic tissues. This novel approach extends the options for gene knock-ins while reducing the risks of insertional mutagenesis, unpredictable expression profiles, and increasing the safety and therapeutic effects.

Introduction

Genomic safe harbors (GSH) may be defined 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 virus vectors, e.g., lentivirus or gammaretrovirus, insert frequently in transcriptionally more active and accessible random euchromatin regions. This indiscriminate integration generates a heterogeneous population of transduced cells with integration events that can result in genotoxicity, gene silencing, unpredictable transgene expression or stability, loss of '"stemness", or dysregulated cellular proliferation, e.g., leukemias. In contrast, targeted integration into a validated GSH site would preclude or minimize these risks. Thus, identification of GSHs that provide safe and reliable locations for gene addition should be a high priority for gene and cellular therapies. Clinical outcomes for lentivirus-dependent therapies addressing blood disorders (multiple myeloma[1], β-hemoglobinopathies[2], coagulopathies[3], or Fanconi anemia[4]), would benefit enormously from an available catalog of verified GSH sites that reduce the risks associated with cell engineering. Importantly, the accessibility of primary human CD34+ hematopoietic stem and progenitor cells[5,6] (HSPCs) combined with programmable endonuclease technologies (particularly CRISPR/Cas9[7]) constitute an exceptional model to assess novel GSH candidates for autologous therapies.

To date, efforts are ongoing to find new human GSH sites beyond the widely applied *AAVS1*[7-9], *CCR5*[10,11], and *hRosa26*[12-14] loci. Proposed criteria identify potential GSH sites that are intergenic and preferentially 50-300 kb distant from known or predicted transcriptional units[15-

[17]. However, the number of characterized sites identified following this approach have been limited and mostly characterized in established cell lines, e.g., HEK239 or K562, with few examples on primary human cells[18,19]. Other criteria based on the use of endogenous mobile elements[20] or chromatin architecture analysis[19,21] have not been particularly effective. First reported to function as GSH by Smith *et al*[9], *AAVS1* remains a popular site for directed transgenesis given that gene knock-in does not affect processes like cellular viability or differentiation[22,23]. However, some reports argue that transgene silencing from *AAVS1* is cell lineage[24] and promoter-specific[25]. The site was originally described as an AAV provirus in exon 1 of *PPP1R12C* in chromosome 19, within a motif consisting of a AAV Rep protein-dependent origin of replication[7,8,26], however, commonly available editing reagents relocated the GSH site to the first intron of *PPP1R12C*[27]. Given that *AAVS1* is a parvovirus integration site, it is highly probably that other parvoviral integration loci may predict GSH sites too.

Parvoviruses are one of the small number of viruses known to have contributed DNA to the vertebrate germline via viral integration and horizontal gene transfer[28]. As a result, endogenous parvoviral elements (EPVs) have resulted from ancient infections of germline cells and persisted during multiple speciation events across geologic time scale[29]. Phylogenetic analysis of EPVs in vertebrate genomes shows that many belong to the extant parvovirus genera *Amdoparvovirus*[30], *Protoparvovirus*[31,32], and *Dependoparvovirus*[29,33]. The viral 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[34], notably syncytin - a retrovirus fusion

protein co-opted by placental mammals for trophoblast fusion early in embryogenesis[35]. Thus, identifying human orthologous sites to EPV loci in non-human vertebrate genomes may indicate suitable sites for gene addition.

In this study, we identify 102 potential human GSH sites by investigating a dataset of EPVs into host species[36]. Ortholog loci were categorized as intergenic or intronic according to their location in the host species and in the human genome. Additionally, the accessibility for genome editing in hematopoietic precursors was approximated using available ATAC-seq data. Through CRISPR/Cas9-mediated genome editing seventeen of these loci (8 intergenic and 9 intronic) were characterized to assess their functionality as GSH sites by inserting an *eGFP* expressing transgene into primary human CD34+ HSPCs. The edited cells displayed no alterations in self-renewal or multipotency, demonstrating suitability of these loci for gene addition. Moreover, transgene expression was maintained across cell differentiation, indicating stable genome integration. To our knowledge this is the first study that provides an extended catalog of human GSH sites with potential applications in CD34+ HSPCs and the results can be extended to other tissues as additional GSH sites 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, 40

intronic) we mapped the orthologous locus in the human genome by utilizing genome-genome alignments and gene synteny[37] 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 precision of the orthologous locations was determined based on sequence identity and gaps in the human to host-species genome alignments.

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 locus 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. The proximity to open chromatin regions was considered using available ATAC-seq data for human CD34+ HSPCs[38], with sites within 2 kb from an ATAC-seq peak being particularly monitored for evaluation. Seventeen loci, 8 intergenic and 9 intronic, were selected for a 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 widely cited *AAVS1* gRNA for comparison purposes[39] (Fig. 1c, Supplementary Table 3). The indel frequencies revealed reasonable propensity to genome editing from all the GSH candidates relative to that for *AAVS1* (Fig. 1c), therefore, we proceeded with a proof-of-concept study targeting primary CD34+ HSPCs derived from the bone marrow of adult human healthy 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)[18,40], or flanked with *gRNA* recognition sites to use a homology-independent targeted integration (HITI)[41]. However, these dsDNA templates are transcriptionally competent as episomal DNA but convey more cytotoxicity by the transfer of naked DNA. Recombinant AAV6 vectors have been engineered according to the mentioned repair pathways showing to improve the gene transfer outcomes[42,43] but they tend to remain episomal after the transduction. Alternatively, short single stranded DNA (ssDNA) templates (<200 nt) were reported to correct point mutations associated with blood disorders[44,45]. Our preliminary comparisons of dsDNA and ssDNA templates displayed reduced GFP background with the latter in absence of CRISPR/Cas9

system (Supplementary Fig. 1A and B). This suggests that the use of ssDNA templates reduces the frequency of random integration of templates that remain episomal after the transfection (Supplementary Fig. 1C and D). Consequently, we engineered ssDNA templates per targeted locus composed of *eGFP* under control of MND synthetic promoter and rabbit polyadenylation signal. The cassettes were flanked by 300 nt homology arms yielding sets of 2 kb templates (Fig. 2a).

Thus, synthetic gRNAs complexed with SpCas9 were co-delivered with 3 μg of the appropriate ssDNA template into 2.5×10^5 primary human CD34+ HSPCs using a nucleofection protocol (Lonza Inc., Morristown, NJ, USA). To minimize transcriptional interference, intronic GSH loci templates were engineered to insert the transgene in the 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 medium for additional 3 days (Fig. 2b). Unless otherwise mentioned, the experiments with undifferentiated CD34+ HSPCs spanned seven days given that prolonged cultures impairs the stemness features of the cells[46]. 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 editing efficiencies in CD34+ HSPCs differed substantially from the values estimated in human embryonic kidney (HEK) 293T cells (Fig. 1c), due to multiple factors including the HEK293T suppressed innate immune responses, overall robustness of this cell line, and hetero-/eu-chromatization differences between tissues. For example, the candidate GSH site *Dep.36* (intergenic) displayed high propensity to 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[38] as an additional criterion to select the best GSH candidates, no correlation was observed between the editing efficiency and the proximity to open chromatin spots (Supplementary Tables 2 and 3). For instance, the targeted *Dep.2* intronic locus produced the highest level of GFP+ cells at two days post-nucleofection (3.23% ± 0.86, P = 0.8938) while being 15 kb from the nearest ATAC peak (Supplementary Table 3). Despite this disparity, persistent GFP expression was monitored through differentiated stages of edited *Dep.2* cells, a phenomenon replicated with other loci not necessarily proximal to ATACseq peaks, particularly 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).

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 was significantly decreased within the three days following the sorting of GFP+ cells from most edited groups compared to control cells (mean ± s.d., 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 cell colonies representing erythroid and myeloid lineages formed independently of the edited GSH locus demonstrating that neither genome modification nor 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 cell population, the long-term hematopoietic stem cells (LT-HSCs) are the cell type with self-renewal and long term engraftment potential representing the most clinically relevant for autologous cell therapies[47]. 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 LT-HSCs were targeted in culture and survived gene addition we analyzed the GFP+ population to distinguish potential LT-HSCs (CD34+/CD90+) from the most abundant progenitor cells (CD34+/CD90-/CD38+, CD34+/CD90-/CD38-)[48] (Fig. 2f). Interestingly, the results suggest that not all loci are accessible in potential LT-HSPC populations. The loci that produced edited LT-HSPCs include *AAVS1*, *Ap.102*, *Dep.3*, *Dep.28*, *Dep.22*, *Prot.181*, *Dep.2*, *Prot.176*, *Dep.1*, *Prot.2*, *Dep.35*, *Dep.55*, and *Prot.218*, with 10.32, 35.97, 12.07, 2.05, 8.56, 10.59, 8.33, 16.67, 34.7, 11.18, 13.22, 26.09, and 5.6%, respectively, making these attractive candidates for long term animal studies.

GSH targeted integration in human CD34+ HSPCs

To verify the integration into the GSH loci of GFP+ HSPCs, genomic DNA was extracted from 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). Considering two GSH alleles per diploid genome, a 50% signal represents one hundred percent of monoallelic integration whereas over 50% indicates (partial) bi-allelic integration. Thus, the editing of *AAVS1* (the reference locus) produced 40% of bi-allelic and 60% of monoallelic 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*, *Dep.55*, and *Prot.218*) resulted uniquely in monoallelic integration likely resulting from a 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 genome editing of *Dep.34*, *Dep.36*, *Dep.28*, and *Dep.56* loci (Fig. 1c) also resulted in undetectable integration events. Factors accounting for this outcome might include the 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) or disruption of vital pathways , thus they are poor candidates for future *in vivo* transplantation studies.

A recent study on gene therapy for patients affected with Wiscott-Aldrich syndrome or β-hemoglobinopathies estimated that as few as 98 corrected HSPCs per 10^6 infused CD34+ cells were able to engraft the bone marrow for long term-production of granulocytes according to the tracking of unique vector insertion sites (IS)[47]. Given that our experiments were designed on a small scale of 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[43,49].

Alternatively, the use of recombinant AAV6 vectors reportedly substantially increases editing efficiencies. To test the latter approach, we used a combined nucleofection/rAAV6 transduction protocol 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 efficiency with the use of ssDNA templates (Fig. 2c and Fig. 2i). As expected, the percentage of sorted GFP+ cells substantially increased two days after transduction without affecting the cell viability (Fig. 2i). Notwithstanding, because the AAV transgenes might persist episomal after internalization, determining the targeted integration measured on sorted cells requires further analysis (Fig, 2h and Supplementary Fig. 1E)

Transcriptional changes in edited CD34+ HSPCs

To determine whether and to what extent the GSH editing perturbs the gene expression of the progenitor cells, RNAseq analysis 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). 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 2163, 925 and 1,144 DEGs, respectively (Figure 3a). Interestingly, the genome editing of intergenic GSH sites *Dep.3* and *Ap.102* resulted with lower numbers of DEGs reporting 509 and 496, respectively (Fig. 3a). Thus, even though the CD34+ HSPCs constitute a heterogeneous cellular population the results suggest that the

targeting of intergenic GSH sites tend 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 = 2.4x10^{-5}) whereas a downstream transcript (*DEPTOR*) decreased by 0.18-fold (P = 2x10^{-11}) (Fig. 3d and Supplementary Fig. 3). Similarly, the *Dep.3* locus downstream transcript *GPNMB* displayed a fair 0.34-fold decrease regarding unedited control cells (P = 0.03) (Fig. 3d and Supplementary Fig. 3).

The gene ontology (GO) analysis ranked more than one group of annotated genes to regulation of the cell proliferation within the top genetic features of the transcriptomes (GO:0008285, GO:0042127, GO:0001558, GO:0008284, and GO:0051726) (Fig. 3c). Notably, the genes associated with malignant transformation showed no evidence of alteration. The evaluated transcriptomes revealed upregulation of tumor suppressor genes commonly *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, leukemia-associated proto-oncogenes displayed downregulation of key transcripts 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. 3). 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 up-

regulation of *HES7* (8.9-fold $P = 0.017$ for *Dep.2*) both transcriptional repressors for the maintenance of blood precursors. In addition, increased transcription of *FRZB* (74.2-fold change, $P < 0.01$), an inhibitor of the β-catenin pathway, was observed (Fig. 3b and 3d). Similarly, all the edited GSH groups displayed up-regulation of genes associated with immune response suggesting cellular responsiveness to *in vitro* manipulation, i.e, nucleofection of DNA and 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 to malignant transformation at the moment of evaluation, although a depth transcriptome analysis with homogeneous cellular systems as well as cell transformation assays remain to be conducted to thoroughly clarify the cellular response to the editing of the novel GSH sites.

Edited CD34+ HSPCs retained multipotency with constitutive or lineage-specific transgene expression

To determine whether gene addition and transgene expression affected 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, GFP expressing colonies derived from edited loci were clearly identified. Edited *Dep.2*, *Dep.3*, *Dep.55*, *Prot.218*, *Prot.2* and *Prot.181* GFP+ colonies from erythroid (BFU-E), granulocyte (CFU-G), granulocyte-macrophage (CFU-GM) and macrophage (CFU-M) lineages could be easily identified as was done for the reference group *AAVS1* (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 (Fig. 2g and Fig.4a) indicating that downregulation occurred on these sites alongside differentiation. Other targeted groups displayed myeloid lineage biased transgene expression based on visual inspection of colonies (*Dep.13*, *Dep.1*, *Ap.102* and *Prot.176*) (Fig.4a). The total colony distributions were similar among the groups as compared to unedited control cells (Fig. 4b). Interestingly, the edited GSH cells with higher fluorescent signal and persistent *eGFP* expression after 14 days of differentiation were those derived from *AAVS1* (control), *Dep.3*, *Dep.2*, *Dep.55*, and *Prot.218* loci (Fig. 4c) being the same with widespread expression (Fig. 4a) and targeting of LT-HSCs (Fig. 2f) becoming thus promising sites for gene addition in human CD34+ HSPCs.

To confirm the multipotency of edited HSPCs, we sorted the erythroid (CD235a+) and myeloid (CD69+) phenotypes within the GFP+ fraction (Fig. 4c and 4d). The composition was similar for most of the analyzed groups relative to the *AAVS1* reference locus (Fig. 4d and 4e). Similar proportions of erythroid and myeloid lineage phenotypes were detected with small portions of unidentified GFP or GFP/CD235a/CD69 triple labeled cells in all the cases (Fig. 4e). Interestingly, the preference for myeloid transgene expression from *Ap.102* and *Dep.1* loci was confirmed by the analysis of cell surface markers indicating tissue specific regulatory mechanisms on the transgene in both loci (Fig. 4e). To ascertain whether epigenetic mechanisms modulated lineage-specific expression, we performed cleavage under target and tagmentation (CUT&Tag) analysis on sorted erythroid and myeloid cells. The results showed enrichment of the silencing mark H3K27me3 on the transgene in erythroid cells derived from *Dep.1* targeted cells as compared to myeloid population. However, an opposite effect was observed between erythroid

and myeloid cells derived from the editing of the intergenic GSH site *Ap.102* where the H3K27me3 mark was more abundant on the transgene in myeloid cells (Fig. 4d). These results are consistent with the analysis of sorted erythroid and myeloid cells.

Persistent transgene expression in terminally differentiated erythroid cells

Given that the final step of erythroid differentiation comprises the erythroblast enucleation, we selected *Dep.2*, *Dep.3*, *Dep.55*, *Prot.218*, *Prot.2* and *Prot.181* groups for further erythrocyte maturation experiments by exposing to human erythropoietin (hEPO) (Fig. 4g). Following the nucleofection, the CD34+ HSPCs were expanded during 14 days 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 those unexposed to hEPO (Fig. 4h). Of note, persistent transgene expression was observed in reticulocytes derived from all analyzed loci except for *Prot.218* (Fig. 4h) as these cells displayed transgene silencing after the exposure to hEPO (Fig. 4f).

Overall, these results support our approach for identifying novel GSH sites in the human genome contributing to establish a concise catalog of verified 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 approximately 200 unique EPV insertions identified[36], 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 HSPC genome editing based on bioinformatic analyses which (i) determined precision of the EPV ortholog in the human genome (Fig. 1a, 1b, 1c, 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 assumptions considered to reduce risks of insertional mutagenesis[15,19], the methodology applied in the current study was intended to identify and characterize lineage-specific GSH loci. The criteria for establishing an ontogenic GSH 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 this approach, we evaluated and characterized novel genomic loci as potential GSHs for gene therapy of blood disorders using primary human HSPCs. However, further evaluation with *in vivo* models is necessary to demonstrate the ability of modified HSPCs to establish long term engraftment of the bone marrow and reconstitution of the blood cell compartments, ideally using a non-human primate model. As a first look, the in vitro results indicate successful editing among the rare and engraftable LT-HSC population (CD34+, CD90+) even using a relatively small number of HSPCs during transfection. For these experiments, the number of nucleofected cells ranged from $2.5x10^5$ to $1.0x10^6$ targeting a single GSH site. 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 2.0-3.0x10^7 autologous CD34+ HSPCs, retrovirally or lentivirally transduced, per kg of body weight as dose for infusion which could mean several rounds of apheresis to achieve the minimal number of cells for manufacturing.

Contrary to the large number of chimeric antigen receptor (CAR) T-cells reportedly required for cancer clinical trials (up to 1.0x10^8 cells/kg)[1,19,50,51], the expanding of human CD34+ HSPCs by prolonged cell culture (>7 days) detrimentally affects the long-term progenitor cells repopulating capacity. Recent studies in long-term transplanted patients of Wiskott-Aldrich syndrome and β-hemoglobinopathies estimated that as few as 98 edited HSPCs per 10^6 infused CD34+ cells correspond to engrafting cells producing gene-corrected granulocytes[47]. This suggests that the more efficient the editing pipeline (which is partially dependent on efficient access to the targeted site), the greater the bioavailability of modified cells in the infusion. Because of the performance of our candidates GSH loci are comparable to the widely applied GSH site, *AAVS1*, the scenario hints that up-scaling for clinical readiness (usually >5x10^6 electroporated cells) might result in increased numbers of engraftable GSH-edited cells for therapeutic applications (i.e., modified LT-HSCs).

Although the GFP+ percentages of sorted cells were low for some targeted GSH loci (e.g. *Dep.23*, *Dep.0, Dep.62*, and *Dep.1*) the values may be easily 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 it has been considered elsewhere[19,50,52]. In this regard, a naturally

occuring insulator in *AAVS1*[53] might be the cause of the robust expression observed from this GSH site (Fig. 2c and 4d)[54] suggesting that insulators artificially incorporated may improve the performance of novel GSH sites. Whether the novel GSH sites contain insulators is unknown, and the ssDNA templates for genome editing used in this report were devoid of such elements. Nevertheless, constitutive reporter expression was maintained as in immature as in both partially and terminally differentiated cells indicating that selected GSH loci might be compatible with the long-term transgene expression in blood cells. Thus, these results establish a catalog of GSH with opportunities to engineer complex transgenes (not limited by size) exploiting the permissively 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.

In conclusion, this study introduces a methodology for locating potential GSH sites in the human genome with a sample of 17 loci yielding at least 9 GSH sites with potential application in the treatment of blood disorders. Moreover, the naturally occurring regulation showed from some of the novel GSH sites can be exploited to get lineage restricted transgene expression, and additional GSH sites are likely to emerge from the remaining mapped loci increasing the options for gene addition approaches. Similarly, these findings can be extended to hunt GSH sites appropriate for other tissues and applications which also increase the possibilities of added traits to modified cells.

Methods

Homology ascertainment relative to human genome

To map the EPV positions to the human genome assembly (GRCh38.p13) we utilized Ensembl for comparative genomic alignments (Ensembl 2023). However, host species genome versions in Ensembl may differ from the EPV loci reports. We verified the location of EPV loci within host genomes through BLASTN or BLAT within Ensembl [33]. Then, using the Comparative Genomics options with Ensembl to examine alignments between the host genome and multiple mammalian genomes to ascertain overall synteny and collinearity, 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 on a closely related host species genome. Gaps in alignment surrounding the EPV element introduce 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 located with a high degree of confidence on the human genome with ambiguities ≤2,000 base pairs (bp). As the experimental portion of this effort was editing human primary CD34+ HSPCs, we examined the accessibility of chromatin of this cell type. HSPC data were obtained from the Gene Expression Omnibus under accession GSE74912. Analysis of Corces et al. ATACseq and RNAseq data was conducted with DolphinNext. Chromatin accessibility was determined through peak calling with MACS2 and special attention was put on those loci within 2 kb of an ATACseq peak.

Cell culture and lipofection

Human embryonic (HEK)-293T cells 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, 5% CO2 and saturating humidity. Transfection experiments on HEK-293T cells were performed in 6-well plates (Corning, approximately 3.5 cm^2). 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 from the cells with DNeasy Blood & Tissue Kit (Qiagen), and following concentration measurement, an aliquot of gDNA was used for PCR amplification to determine editing efficiency and specificity. Human bone marrow CD34+ stem and progenitor cells (HSPCs), obtained from healthy adult donors, were purchased from Stem Cell Technologies (SCT, Cambridge, MA 02142). Briefly, thawed cells were pre-stimulated for 48 h in StemSpanTM SFEM II medium (SCT #09655) with Expansion supplement (SCT 02691) and 1 µM UM729 (SCT 72332), at 37° C, 5% CO_2 and saturating humidity. Following electroporation, the CD34+ HSPCs were returned to the same culture conditions for additional 72 h before further downstream use.

Guide RNAs for CRISPR/Cas9 gene editing

Single guide RNAs (gRNAs) were engineered by simultaneous use of CHOPCHOP, CRISPOR and CRISTA designing tools. Three to five highly scored gRNAs, with minimal off-target sites, were chosen per candidate GHS locus. Then, oligonucleotides coding for *gRNA* sequences were *in vitro*

aligned and cloned into pX330 plasmid (Addgene, Watertown, MA 02472) for co-expression of *Cas9* and a corresponding *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 TIDE analysis tool. 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) to be 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. Briefly, an expression cassette coding for eGFP under control of 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. 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). DpnI treatment (NEB R0176) was performed prior to the ssDNA synthesis to eliminate the plasmid templates from the final preparations.

Electroporation

Prior to nucleofection of CD34+ HSPCs, ArchiTect™ Cas9 nuclease (SCT) and the corresponding sgRNA were mixed (1:2.5 molar ratio) to form a ribonucleoprotein complex at 25° C. Assembled complexes were held at 4° C until nucleofection. In parallel, 2 day pre-stimulated CD34+ HSPCs were washed in PBS before being resuspended in P3 nucleofector solution (Lonza, V4XP-3032). An equivalent of $2.5x10^5$ CD34+ HSPCs were mixed with gRNA:Cas9 complexes and 3 µg ssDNA donor template (20 µL) in nucleofector strip format (Lonza). Electroporation was performed in Unit X, Lonza 4D nucleofector, using the DZ-100 program. After nucleofection, cells were incubated for 10 min at room temperature before adding the StemSpanTM SFEM II medium with Expansion supplement and 1 µM UM729 for a 2 day-recovery time. After the recovery period, the cells were sorted by flow cytometry to collect the GFP+ populations.

AAV6 transduction

CD34+ HSPCs were thawed in StemSpan™ SFEM II medium for a 2 days pre-stimulation period under conditions described above. Then, $2.5x10^5$ CD34+ cells were electroporated in presence of Cas9:gRNA complexes, followed by a recovery period of 15 min at RT. Next, cells were collected and transduced with rAAV6s at a multiplicity of infection (MOI) of 200,000-300,000 gc/cell in a Falcon 15 mL conical tube (Corning). Cells were transferred to a 6-well culture plate (Corning) at

a density of $1x10^5$ cells/mL for 2 days before downstream analysis. AAV6 vectors were produced by plasmid transfection of HEK293 cells according to established protocols of Viral Vector Core, Gene Therapy Center, UMASS Medical School.

Flow cytometry

After 2 days of recovery time, nucleofected CD34+ HSPCs were washed once in PBS and resuspended in ice cold 1% FBS phenol free-IMDM (Gibco) to be processed by fluorescence activated cell sorting (FACS) using non-transfected samples to set the GFP+ gate. Singlets were gated on the scatter plot. Propidium iodide (Sigma) was included for viability determinations then, cells were collected and cultured for 3 additional days in StemSpan™ SFEM II, CD34+ Expansion supplement and 1 µM UM729. The fold expansion of the cells 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 experiments with 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 1% FCS PBS. Subsequently, cells were incubated in presence of TruStain FcX™ (BioLegend) before staining with antibodies against CD34, CD90 and CD38. Samples were fixed in a 2% PFA solution washed and resuspended in phenol free-IMDM. The GFP+/CD34+ population was gated to explore the modified immature cell populations. For cell lineage determinations (erythroid and myeloid), cells were recovered by diluting the methylcellulose in PBS and resuspending in ice cold 1% FCS in PBS. 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.

Colony Forming Unit assay (CFU)

Between 300 and 1500 sorted GFP+ cells were seeded in 1 mL methylcellulose medium (SCT H4435). Differentiation was induced during 14 days at 37°C, 5% CO_2 and saturating humidity before scouring the Colony Forming Units (CFUs) under bright field of Axiovert 135 microscope (Zeiss). Fluorescent images were obtained with fluorescent microscopy (LionHeart FX,BioTek). After colony scoring, the 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.

Erythroid differentiation

CD34+ HSPCs nucleofected with the appropriate CRISPR/Cas9-ssDNA editing set, 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 s 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, after which nuclear staining with Hoechst 33342 (ThermoFisher) was performed prior to microscopy imaging.

Droplet digital PCR

To determine the integration percentage among manipulated cell populations, genomic DNA was isolated from the bulk of differentiated cells (Qiagen) and the transgene - genomic DNA junctions were 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. Primer and probe sequences are detailed in Supplementary Tables 5, 6 and 7. Reaction mixtures were set up with 1x ddPCR Supermix, minus dUTP, for probes (Bio-Rad), primer/probe mixtures (900 nM primer/250 nM probe) and up to 60 ng gDNA. The following temperature settings was employed for the thermocycler: initial denaturation for 95°C for 10 min with ramp rank of 2°C/sec, followed by 60 cycles of denaturation at 94°C for 30 sec (ramp rank of 2°C/sec), annealing at 55°C for 1 min (ramp rank of 1°C/sec), and extension at 72°C for 2 min (ramp rank of 1°C/sec), and a terminal denaturation step at 98°C for 10 min, followed by cooling the reaction to 4°C. Droplet Reader and Quantasoft (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

GFP+ HSPCs were sorted using flow cytometry and cultured for three days post-sorting in StemSpanTM SFEM II medium (SCT) before being resuspended 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* GRCh38 reference genome available on ENSEMBL using STAR aligner v.2.5.2b. Differential expression analysis was performed by providing un-normalized read counts to DESeq2 with donor as a cofactor. Genes with an adjusted p-value < 0.05 and absolute $\log_2$ fold change > 1 were called as differentially expressed genes. 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).

Statistical analysis

All the experiments with primary human cells were independently performed at least three times for statistical analysis with different 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

Raw RNAseq data are available from the Sequence Read Archive under BioProject Accession XXXYYY.

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).

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.

References

1 Stadtmauer, E. A. *et al.* CRISPR-engineered T cells in patients with refractory cancer. *Science* **367**, doi:10.1126/science.aba7365 (2020).

2 Drysdale, C. M. *et al.* Hematopoietic-Stem-Cell-Targeted Gene-Addition and Gene-Editing Strategies for beta-hemoglobinopathies. *Cell Stem Cell* **28**, 191-208, doi:10.1016/j.stem.2021.01.001 (2021).

3 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).

4 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).

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 Pavel-Dinu, M. *et al.* Gene correction for SCID-X1 in long-term hematopoietic stem cells. *Nat Commun* **10**, 1634, doi:10.1038/s41467-019-09614-y (2019).

7 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).

8 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).
9 Smith, 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).
10 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).
11 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).
12 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).
13 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).
14 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).
15 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).
16 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).
17 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).
18 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).
19 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).
20 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).
21 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).
22 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).
23 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).

24 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).
25 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).
26 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).
27 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).
28 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).
29 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).
30 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).
31 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).
32 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).
33 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).
34 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).
35 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).
36 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).
37 Martin, F. J. *et al.* Ensembl 2023. *Nucleic Acids Res* **51**, D933-D941, doi:10.1093/nar/gkac958 (2023).
38 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).
39 Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823-826, doi:10.1126/science.1232033 (2013).
40 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).

41 Kelly, J. J. *et al.* Safe harbor-targeted CRISPR-Cas9 homology-independent targeted integration for multimodality reporter gene-based cell tracking. *Sci Adv* **7**, doi:10.1126/sciadv.abc3791 (2021).
42 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).
43 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).
44 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).
45 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).
46 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).
47 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).
48 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).
49 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).
50 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).
51 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).
52 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).
53 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).
54 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).

Figure legends

Fig. 1. Identification of 102 potential genome safe harbor (GSH) sites in the human genome. **a** Circle plot of EPV orthologues mapped on 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 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 group 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* expressing cassettes into GSH candidates using specific CRISPR/Cas9-ssDNA editing sets. Single-stranded DNA donor templates were flanked with 300 nt homology arms to mediate homologous recombination through single stranded template repair (SSTR) pathway. **b** Experiments with bone marrow CD34+ HSPCs. Pre-stimulated cells were nucleofected in presence of the CRISPR/Cas9-ssDNA editing set and cultured for two days before the sorting of the GFP+ population. The GFP+ cells were cultured for an additional three days before downstream processing. Alternatively, AAV6 vectors were used to deliver templates to target *Ap.102*, *Dep.33*, *Dep.13*, *Dep.1* and *Dep.35* loci. In parallel, 1 day-nucleofected cells were cultured in erythroblast expansion conditions before maturation with human EPO. **c** Percentage of GFP+ cells at two days post-nucleofection. Intergenic and intronic loci are depicted as red and blue bars, respectively. Open circles represent individual human donors. *AAVS1* GSH (gray) served as reference control. Mean ± s.d, ANOVA followed by Dunnett test * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$ and **** $P < 0.0001$. $n \geq 4$ independent experiments. **d** Cell viability determined by flow analysis of propidium iodide stained cells. Comparison control corresponds to unedited cells (mock). Open circles represent independent human donors. Mean ± s.d, ANOVA followed by Dunnette test, $n \geq 3$ independent experiments. **e** Proliferation rate of GFP+ HSPCs three days after sorting. ANOVA followed by Dunnett test, $n \geq 3$ independent experiments. **f** Targeting of long term-HSCs (CD34+/CD90+, magenta) and progenitor cells (CD34+/CD38+, CD34+/CD38-) within the GFP+/CD34+ fraction. n = 2-3 independent experiments using different human donors. **g** Targeted Integration in GFP+ cells sorted after 14 days of liquid culture. Specific primer/probe sets (arrows and bars) were 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.

Fig. 3. **Gene expression changes following the targeting of GSH candidates in primary human HSPCs**. **a** Comparison of total DEGs among intergenic (*Dep.3*, *Ap.102*) and intronic (*Dep.2*, *Dep.55*, *Prot.218*) GSH candidates, red and blue, respectively. The reference GSH *AAVS1* is 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. 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 cultivation in differentiation medium (bright field and epifluorescence). The GSH name is indicated on the top, the category on the bottom. and the CFU on the left (E, erythroid; G, granulocyte; GM, granulocyte–macrophage and M, macrophage). *AAVS1*, *Dep.2*, *Dep.3*, *Dep.55*, *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 myeloid (CD69+) composition of GFP expressing cells. Representative FACS plots are shown (left). **d** Normalized total reads detected on the transgene in sorted myeloid and erythroid populations derived from the targeting of *Dep.1* or *Ap.102* GSH sites. **e** Stages of erythroid maturation. **f** Erythroid maturation from *AAVS1*, *Dep.2*, *Dep.3*, *Prot.181*, *Dep.55, 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.

Figure 1

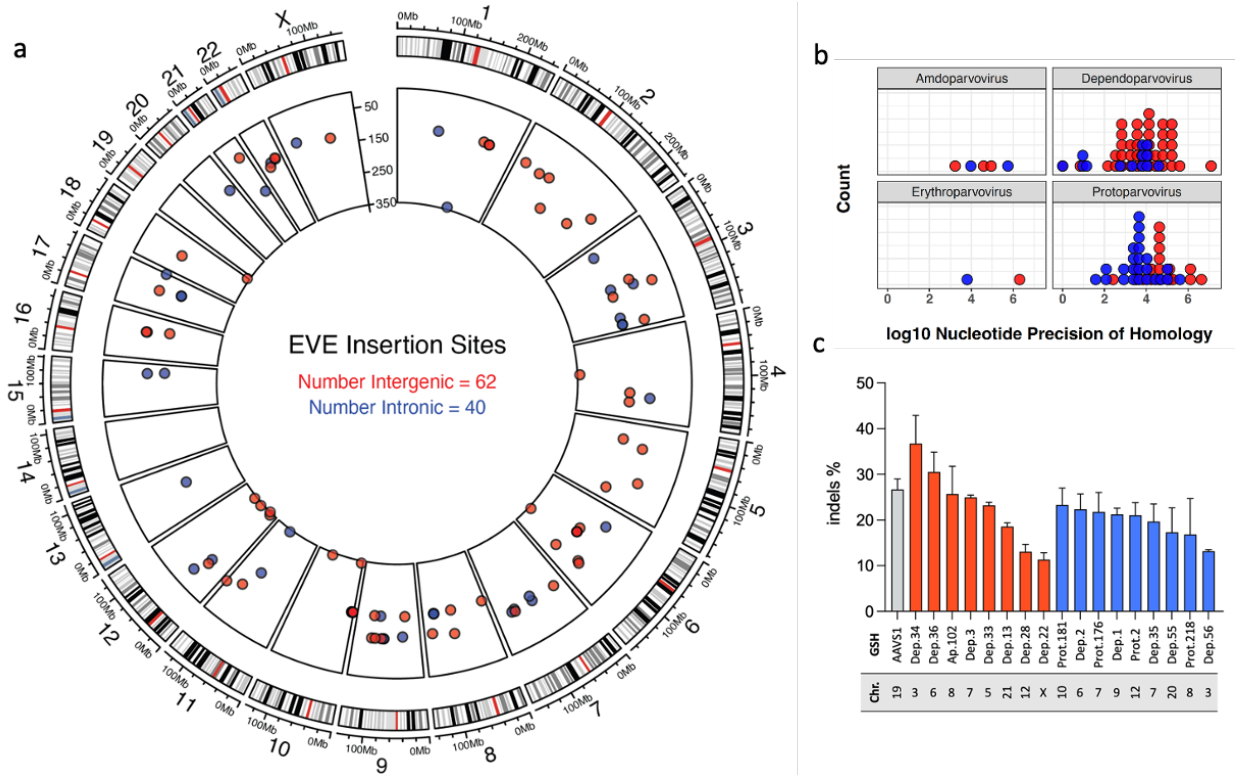

a
EVE Insertion Sites
Number Intergenic = 62
Number Intronic = 40
b
Amdoparvovirus
Dependoparvovirus
Erythroparvovirus
Protoparvovirus
Count
log10 Nucleotide Precision of Homology
c
indels %
GSH
AAVS1
Dep.34
Dep.36
Ap.102
Dep.3
Dep.33
Dep.13
Dep.28
Dep.22
Prot.181
Dep.2
Prot.176
Dep.1
Prot.2
Dep.35
Dep.55
Prot.218
Dep.56
Chr.

Figure 2

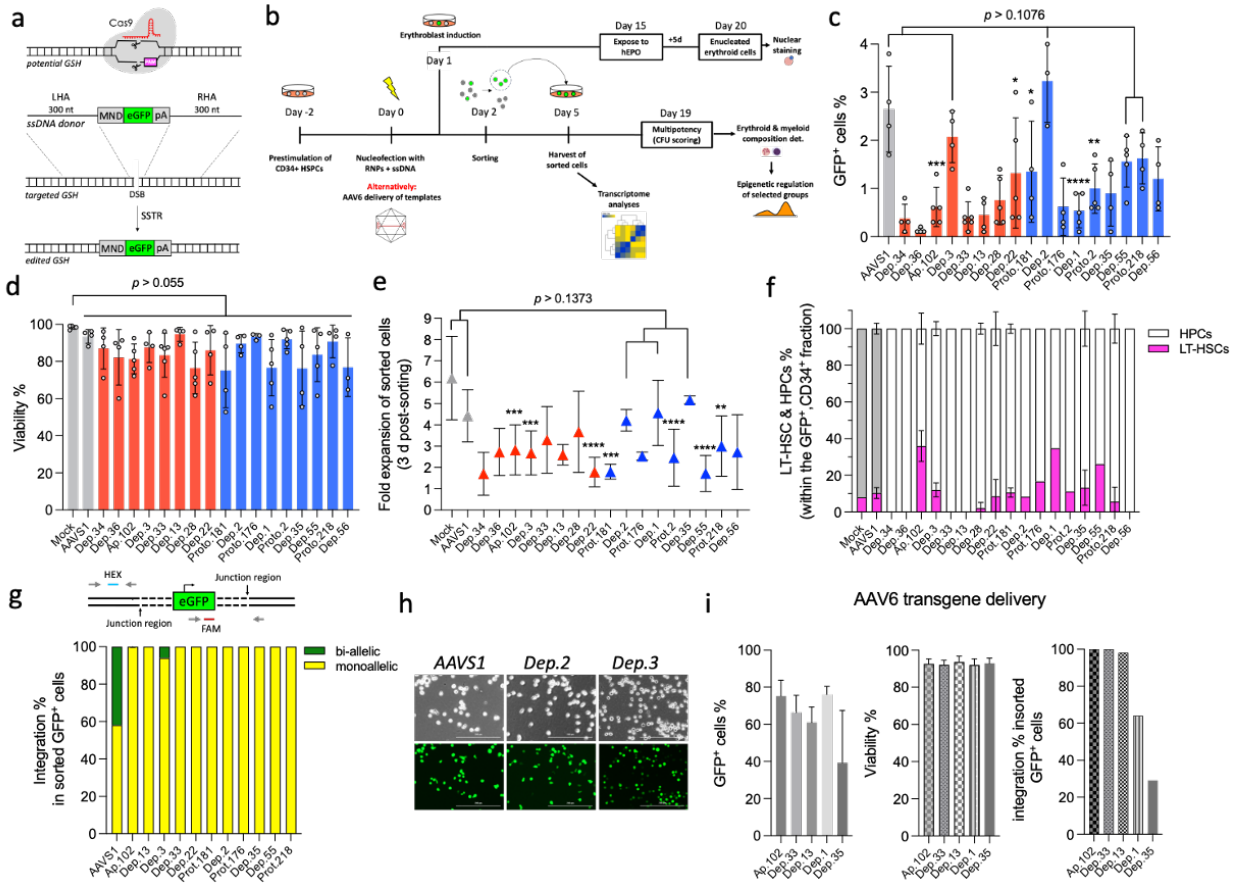

a
Cas9
potential GSH
LHA
300 nt
RHA
300 nt
MND
eGFP
pA
ssDNA donor
targeted GSH
DSB
SSTR
edited GSH
b
Erythroblast induction
Day 1
Day 15
Expose to hEPO
+5d
Day 20
Enucleated erythroid cells
Nuclear staining
Day -2
Day 0
Day 2
Day 5
Day 19
Multipotency (CFU scoring)
Erythroid & myeloid composition det.
Prestimulation of CD34+ HSPCs
Nucleofection with RNPs + ssDNA
Alternatively: AAV6 delivery of templates
Sorting
Harvest of sorted cells
Transcriptome analyses
Epigenetic regulation of selected groups
c
p > 0.1076
GFP+ cells %
d
p > 0.055
Viability %
e
p > 0.1373
Fold expansion of sorted cells (3 d post-sorting)
f
LT-HSC & HPCs % (within the GFP+,CD34+ fraction)
HPCs
LT-HSCs
g
HEX
Junction region
eGFP
FAM
Integration % in sorted GFP+ cells
bi-allelic
monoallelic
h
AAVS1
Dep.2
Dep.3
i
AAV6 transgene delivery
GFP+ cells %
Viability %
integration % insorted GFP+ cells

Figure 3

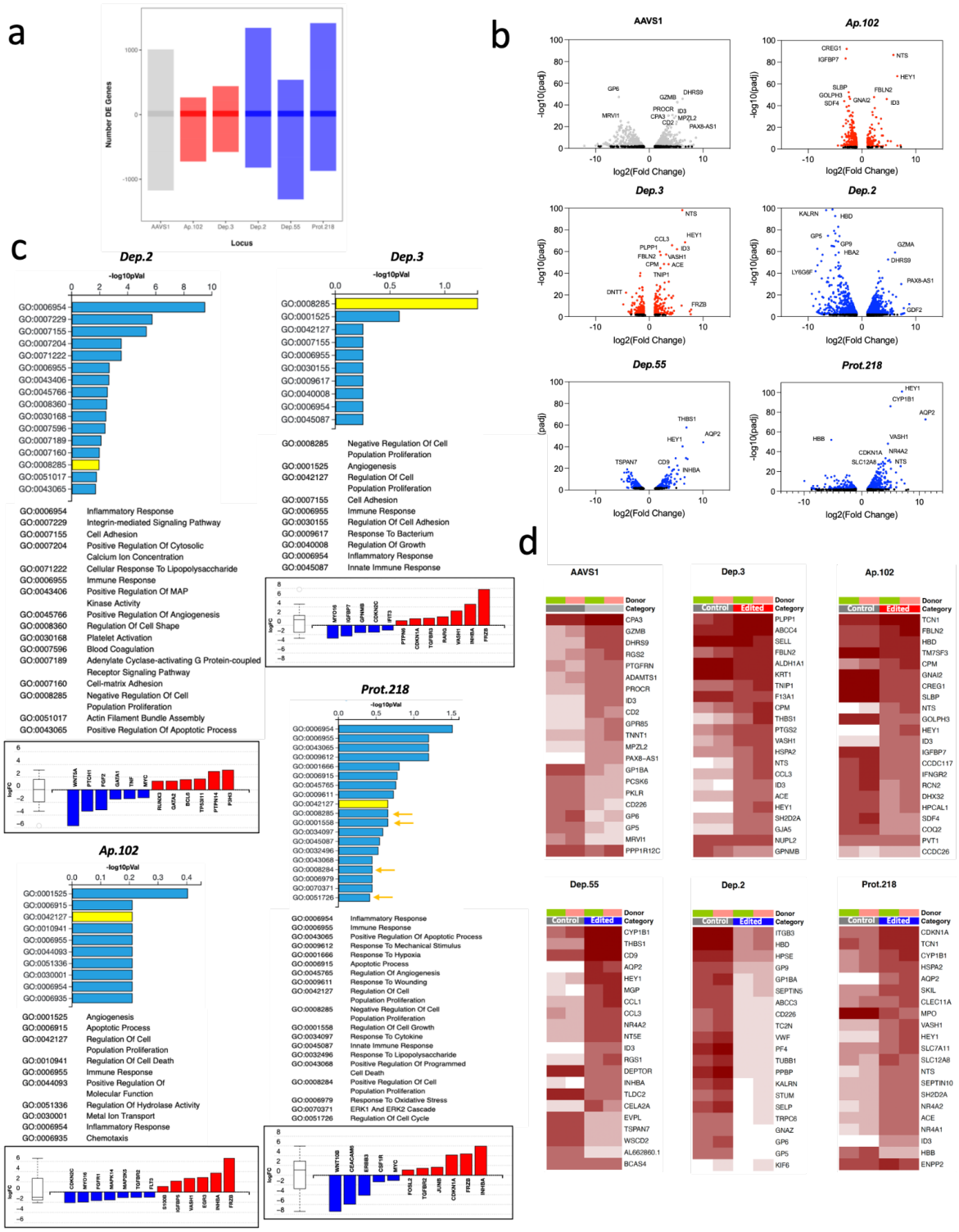
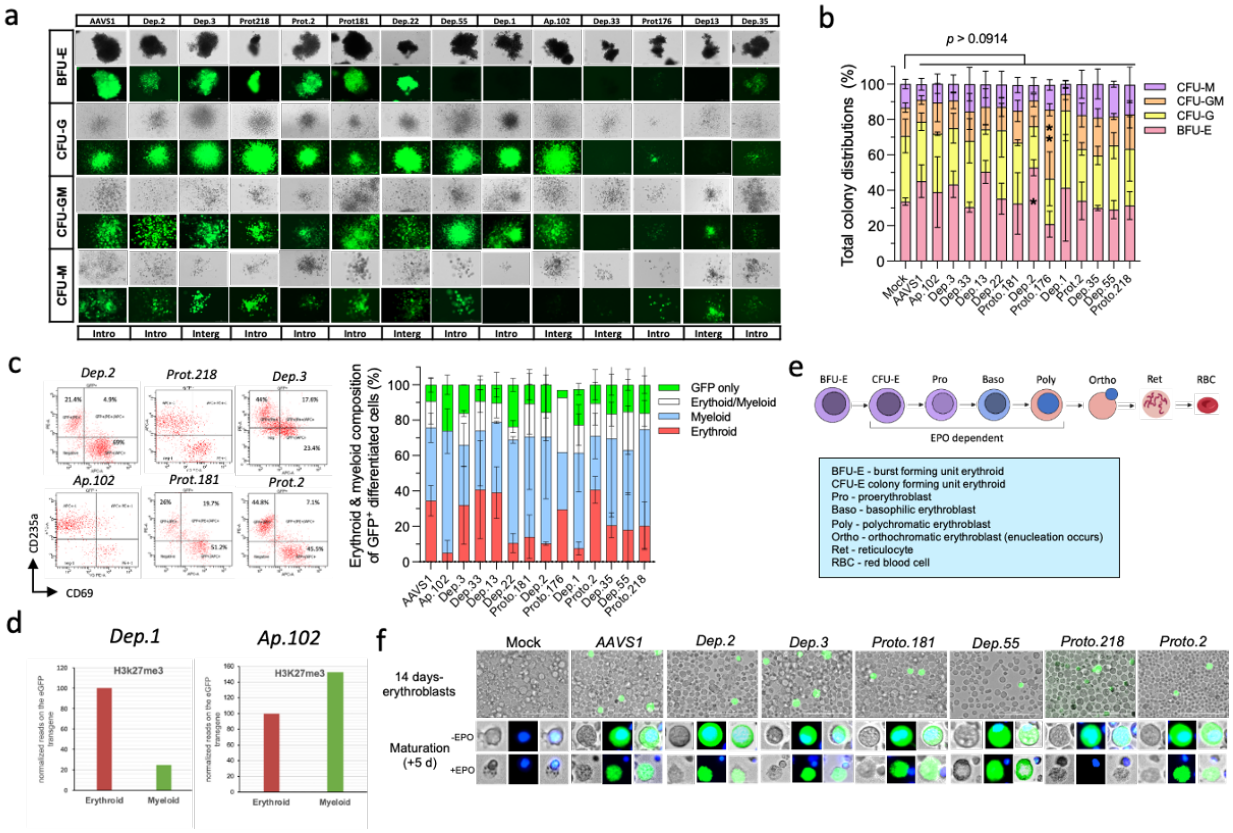

a
Number DE Genes
Locus
AAVS1
Ap.102
Dep.3
Dep.2
Dep.55
Prot.218
b
AAVS1
Ap.102
Dep.3
Dep.2
Dep.55
Prot.218
-log10(padj)
log2(Fold Change)
c
Dep.2
Dep.3
Prot.218
Ap.102
-log10pVal
GO:0006954 Inflammatory Response
GO:0007229 Integrin-mediated Signaling Pathway
GO:0007155 Cell Adhesion
GO:0007204 Positive 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
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
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
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
logFC
d
AAVS1
Dep.3
Ap.102
Dep.55
Dep.2
Prot.218
Donor
Category
Control
Edited

Figure 4


a
AAVS1
Dep.2
Dep.3
Prot218
Prot.2
Prot181
Dep.22
Dep.55
Dep.1
Ap.102
Dep.33
Prot176
Dep13
Dep.35
BFU-E
CFU-G
CFU-GM
CFU-M
Intro
Interg
b
p > 0.0914
Total colony distributions (%)
CFU-M
CFU-GM
CFU-G
BFU-E
Mock
AAVS1
Ap.102
Dep.3
Dep.33
Dep.13
Dep.22
Proto.181
Dep.2
Proto.176
Dep.1
Proto.2
Dep.35
Dep.55
Proto.218
c
Dep.2
Prot.218
Dep.3
Ap.102
Prot.181
Prot.2
CD235a
CD69
Erythroid & myeloid composition of GFP+ differentiated cells (%)
GFP only
Erythoid/Myeloid
Myeloid
Erythroid
e
BFU-E
CFU-E
Pro
Baso
Poly
Ortho
Ret
RBC
EPO dependent
BFU-E - burst forming unit erythroid
CFU-E colony forming unit erythroid
Pro - proerythroblast
Baso - basophilic erythroblast
Poly - polychromatic erythroblast
Ortho - orthochromatic erythroblast (enucleation occurs)
Ret - reticulocyte
RBC - red blood cell
d
Dep.1
Ap.102
H3k27me3
Erythroid
Myeloid
f
Mock
AAVS1
Dep.2
Dep.3
Proto.181
Dep.55
Proto.218
Proto.2
14 days-erythroblasts
Maturation (+5 d)
-EPO
+EPO