HLF Expression Defines the Human Haematopoietic Stem Cell State

Bernhard Lehnertz[1, #], Tara MacRae[1], Jalila Chagraoui[1], Elisa Tomellini[1], Sophie Corneau[1], Nadine Mayotte[1], Isabel Boivin[1], and Guy Sauvageau[1, 2, 3, #]

[1]Molecular Genetics of Stem Cells Laboratory, Institute for Research in Immunology and Cancer, University of Montreal, Montreal QC, H3T 1J4, Canada

[2]Division of Hematology, Maisonneuve-Rosemont Hospital, Montreal QC, H1T 2M4, Canada

[3]Department of Medicine, Faculty of Medicine, University of Montreal, Montreal QC, H3T 1J4, Canada

[#]correspondence: Bernhard Lehnertz (bernhard.j.lehnertz@umontreal.ca), Guy Sauvageau (guy.sauvageau@umontreal.ca)

Abstract:

Haematopoietic stem cells (HSCs) sustain blood cell homeostasis throughout life and are able to regenerate all blood lineages following transplantation.

Despite this clear functional definition, the accurate separation of human HSCs from their differentiated progeny can currently only be achieved through combinatorial measurement of multiple surface antigens.

Here, we identify the master transcription factor *HLF* (*Hepatic Leukemia Factor*) as the most selective and consistently expressed marker gene in HSC-enriched sub-populations of different ontogenetic sources and *ex vivo* expanded CD34^{+} cord blood cells. To directly track its expression, we have developed a genomic *HLF* reporter strategy, capable of selectively labeling the most immature human blood cells on the basis of a single engineered parameter.

Most importantly, *HLF*-expressing cells comprise all of the stem cell activity in culture and *in vivo* during serial transplantation.

Taken together, these results establish *HLF* as the most defining gene of the human haematopoietic stem cell state and outline a next-generation approach to continuously mark these cells with unmatched fidelity.

Introduction:

The existence of rare, serially transplantable and multipotent haematopoietic stem cells (HSCs) was first demonstrated nearly sixty years ago in mice[1-3]. Since then, HSCs have been studied extensively not only due to their unique biology but because of their paramount regenerative potential that is now widely exploited in the clinic[4].

Despite their high clinical relevance, the exact molecular identification of human HSCs remains unresolved and invariably requires profiling of complex stem-cell associated surface marker combinations[5]. Since most of these surface markers are key mediators of cellular homing or signaling, their expression is often up- or down-regulated in response to changing physiological conditions, significantly impacting their utility during certain experimental procedures, most notably *ex vivo* culture[6,7].

In mice, several HSC-enriched genes encoding intracellular proteins have been identified and, using transgenic reporter strains, were demonstrated to label repopulating cells with accuracies that rival the most sensitive HSC surface marker panels[8-14]. Collectively, these studies suggest that a number of intracellular proteins, particularly transcriptional and chromatin regulators, may be much more specifically expressed than most, if not all, currently used surface HSC markers. This notion supports the idea that, in addition to interaction with the niche, the haematopoietic stem cell state is largely determined on the transcriptional level, likely through combined activity of these factors.

With recent technological advancements in targeted gene editing using CRISPR[15] and in functional expansion of human HSCs in culture[16-20], the use of genetic reporter alleles in these cells has become conceivable without the necessity to modify the human germ line. Moreover, improved characterization of developmental gene expression networks using single cell transcriptomics has set the stage for population marker gene identification with unprecedented resolution. Through integration of these key advancements, we set out to identify the most

selectively expressed candidate genes in human HSCs and engineer a genomic reporter allowing prospective identification of *bona fide* human HSCs in culture and *in vivo*.

Results

Specific Expression of *Hepatic Leukemia Factor* (*HLF*) in enriched human HSC populations.

We previously identified surface markers such as CD201 and ITGA3 (CD49c) that best define long-term repopulating cells in optimized *ex vivo* $CD34^+$ cord blood stem cell expansion conditions[7,21].

Integrated transcriptome analysis of these highly enriched ($CD34^+/CD201^+$ and $CD34^+/CD45RA^{low}/CD201^+/CD90^+/CD133^+/ITGA3^+$) versus depleted LT-HSC populations yielded a set of genes (n=17) with the most significant LT-HSC-associated expression (**Supplemental Fig. E1**). *Hepatic Leukemia Factor* (*HLF*) ranked highest in this list based on fold-change in expression between LT-HSCs and differentiated cells (**Fig. 1a**). Furthermore, *HLF* is not expressed in mature peripheral blood cells[22] (**Fig. 1b**) and single-cell transcriptomes of freshly isolated and culture-expanded $CD34^+$ cord blood cells confirmed highly enriched *HLF* expression in the haematopoietic stem and progenitor cell (HSPC) cluster (**Fig. 1c, pink cluster and Fig. 1d and e**). Moreover, within this cluster, *HLF* expression rapidly decreased as cells progressed towards lineage commitment, a pattern that aligns with latest models of gradual rather than stepwise HSC differentiation[23,24].

Next, we benchmarked the expression profile of *HLF* against (i) HSC-associated genes we and others identified (*AVP*[25], *MLLT3*[26]) or that have previously been characterized in mice (*PRDM16*[27], *GATA3*[8], *HOXB5*[11], *MEIS1*[12], *MECOM*[14], *FGD5*[9] and alpha-Catulin (*CTNNAL1*)[10]), as well as against (ii) surface markers commonly used to prospectively isolate human HSCs (*CD34*[28],

CD201 (*PROCR*)[7], CD49c (*ITGA3*)[21], CD133 (*PROM1*)[29], CD90 (*THY1*)[30] and CD49f (*ITGA6*)[31]) (**Fig. 1f and Supplemental Fig. E2**). While varying degrees of HSPC-enriched expression were detectable for most of these genes (**pink density profiles in Fig. 1f and Supplemental Figure E2, Supplemental Tables E3 and E4**), *HLF*, *AVP*, *GATA3*, *MEIS1*, *HOXB5*, *MLLT3* and *MECOM* displayed the most specific expression within HSPCs of freshly purified human CD34$^+$ cord blood cells (**Fig. 1f, left panel**). Of note, these genes generally performed better than HSC-associated surface antigens, consistent with the requirement to stain for several of these markers in combination to achieve high HSC enrichment.

In seven-day UM171-supplemented cultures, *HLF*, *AVP*, *PRDM16* and *GATA3* exhibited the highest HSPC enrichment (**Fig. 1f, right panel**), suggesting that they may be particularly suitable to mark immature cells under *ex vivo* expansion conditions.

To confirm expression of these genes in different developmental and physiological contexts, and to assess their specificity in a larger diversity of haematopoietic lineages and intermediates, we investigated single cell transcriptomic datasets of adult bone marrow[32] (**Fig. 1g-i and Supplemental Fig. E5a**) and fetal liver[33] (**Fig. 1j-l and Supplemental Fig. E5b**), each aggregating more than 100,000 cells from several bio-informatically integrated specimens. Within haematopoietic cells, representing most cells in both datasets, *HLF* exhibited pronounced HSPC-restricted expression, with the only exception being negligible expression in a small subset of naïve T-cells in adult bone marrow (**Fig. 1h**). Furthermore, *HLF* expression was readily detectable in the stromal fraction of adult bone marrow, as well as in fetal liver fibroblasts and hepatocytes (**Fig. 1g-i**).

Based on these results we conclude that within the blood system, *HLF* is the gene with the most selective expression in HSC-enriched sub-populations across their most relevant anatomical and ontogenetic sources. As such, it can be regarded as the prime candidate to mark human haematopoietic stem cells independent of developmental or environmental (i.e. ex vivo culture) context.

Engineering of a genomic *HLF*-reporter transgene in human cells.

We reasoned that *HLF*-expression, if visualized genetically, could provide an efficient readout to identify immature human blood cells. We therefore devised a strategy to introduce a fluorescent reporter cassette into the endogenous *HLF* locus with an approach that combines nucleofection of a Cas9/sgRNA ribonucleoprotein complex with delivery of a homologous recombination (HR) template by recombinant adeno-associated virus (rAAV6) transduction[34]. As outlined in **Fig. 2a**, we targeted a guide RNA binding site downstream of the *HLF* open reading frame (**Supplemental Fig. E6**) to knock-in an internal ribosome entry site (IRES) followed by an expression cassette encoding ZsGreen (ZsG) and either P2A-linked Puromycin resistance or a truncated EGFR[35], facilitating possible additional modes of selection. This strategy fully retained *HLF* coding potential and its 3' untranslated region expression by ensuring use of its endogenous poly-adenylation signal, important considerations since *HLF* plays vital functions in mouse and human HSCs[36-38].

We first tested this approach in *HLF*-expressing HepG2 and non-expressing HEK293 cells. In both lines, reporter targeting to the *HLF* locus was achieved as detected by droplet digital PCR (ddPCR) (**outlined in Fig. 2c**) with cognate pairing of guide RNA and repair template (**Fig. 2b middle and right panels**) but not when a control guide RNA (sgAAVS1) was used (**Fig. 2b, left panel**). Importantly, targeted reporter integration resulted in stable ZsG expression in HepG2 but not in HEK293 cells (**Fig. 2d**), as expected from endogenous *HLF* expression levels in these cells (**Fig. 2e**). These results not only provided proof-of-principle of the experimental approach but also demonstrated reporter functionality and selectivity.

We then adapted this protocol to cord blood $CD34^+$ cells which, after 3 days of pre-expansion with UM171, were fractionated into HSC-enriched and depleted populations based on high or low CD201 surface expression, respectively[7] (**Fig. 2f and g**). These sub-fractions, as well as unsorted (bulk) cells were expanded for an additional 24h, electroporated with Cas9/sgHLF RNP and

transduced with the repair template encoding rAAV6 over a dilution range of virus-to-cell ratios (multiplicity of infection, MOI100-1000). Finally, after three additional days of HSC-supportive culture, ddPCR and FACS analysis to evaluate HR efficiencies and reporter expression was performed (**Fig. 2f**).

Within corresponding experimental conditions, HR frequencies were similar between bulk, HSC-enriched and depleted fractions and reached a maximum of ~50% with the highest tested rAAV6 virus titer of MOI1000 (**Fig. 2h top panels**). Strikingly, while HSC-depleted CD201- cultures gave rise to no ZsG^+ cells (**Fig. 2h bottom right panel**) despite successful HR, reporter expressing cells were readily detectable in all other rAAV6-containing conditions (**Fig. 2h bottom panels**). Importantly, pre-enrichment of CD201+ cells and rAAV6 transduction at MOI400 yielded the highest number of ZsG^+ cells among all tested conditions (**Fig. 2h centre bottom panel**). Of note, HR reached only intermediate levels with these parameters, suggesting that an excessive rAAV6 vector-to-cell ratio is toxic for *HLF*-expressing cells.

Since Cas9/sgRNA electroporation significantly impacted the overall cell recovery at day seven compared to untreated controls (**Fig. 2h middle panels**), we next tested whether transient p53 inhibition could alleviate some of this loss in cell vitality, particularly within the stem cell compartment, as suggested[39]. Indeed, co-electroporation of synthetic siRNA against p53 resulted in a threefold enhanced recovery of *HLF-ZsG* expressing cells (**Fig. 2i, bottom panel**), while overall cell numbers improved only marginally (**Fig. 2i, middle panel**). Of note, this effect was not explained by increased HR efficiency (**Fig. 2i, top panel**).

In conclusion, our data establishes an experimental approach to visualize endogenous *HLF*-expression through targeted genomic integration of a fluorescent reporter cassette that is feasible not only in human cell lines but more importantly, in cord blood HSPCs.

Selective *HLF*-transgene expression in immunophenotypic human LT-HSCs.

We next evaluated the *HLF-ZsG* reporter expression in the context of stem cell specific surface marker panels adapted for cultured cord blood cells[7,21].

Based on the optimized conditions including co-nucleofection of siTP53 and rAAV6 transduction at MOI400, comparable allelic HR ratios of ~20% were measured in both HSC-depleted and HSC-enriched sub-fractions (**Fig. 3a**, CD201- and CD201$^+$ sorted), corresponding to 20-40% of cells carrying at least one *HLF-ZsG* reporter allele. *HLF-ZsG* transgene expression was restricted to the HSC-enriched fraction, reached a mean of 1.17% (**Fig. 3b**, CD201$^+$ sorted) and yielded similar ZsG$^+$ cell numbers as in preliminary experiments (**Fig. 3c**). Strikingly, reporter-positive cells largely expressed LT-HSC surface phenotypes defined by characteristic combinations of CD34, CD45RA, CD201, CD90 and ITGA3 (**Fig. 3d and e**). Dimensionality reduction of these surface marker and *HLF-ZsG* expression profiles indicated that *HLF-ZsG* expression defines a concise sub-population which clusters inside increasingly restricted immuno-phenotypically defined HSC populations (**Fig. 3f**).

Furthermore, *HLF-ZsG* transgene expression increased within progressively restrictive HSC surface marker combinations (**Fig. 3g and h**). Extrapolation of these results suggested that the overall fraction of *HLF*-expressing cells was between 2.9% and 5.85%, based on extreme scenarios where either all HR positive cells contained exclusively mono-allelic or bi-allelic integration, respectively (**Fig. 3i total**). Accordingly, the most stringent LT-HSC enriching surface combination (CD34$^+$/CD201$^+$/CD90$^+$/ITGA3$^+$) contained around 7.5% *HLF-ZsG* expressing cells (**Fig. 3g and h, red**), which can be extrapolated to at most 40% *HLF*-expression within this sub-fraction (**Fig. 3i**). Taken together, these results suggested that *HLF-ZsG* expression *per se* can surrogate multi-parametric FACS analysis to potentially identify the most immature cells in *ex vivo* cord blood cell cultures.

***HLF* expression identifies repopulating cells in CD34+ cord blood cultures.**

To test the ability of the reporter to identify functional HSCs, we sorted and transplanted either bulk *HLF-ZsG* targeted HSPCs (**Fig. 4a, black colour code**), reporter expressing (**Fig. 4a, green**) or non-expressing sub-populations (**Fig. 4a, blue**). To control for adverse effects of the targeting procedure, two additional cohorts received either untreated parental cells (**Fig. 4a, red**) or cells electroporated with a neutral guide RNA (sgAAVS1) (**Fig. 4a, purple**). Importantly, transplanted cell doses were corrected to reflect the ratio of reporter positive versus negative subsets in the total targeted population (**summarized in Fig. 4b**).

Next, we assessed the engraftment potential of these cells in transplanted NSGS recipients at short (3 weeks), intermediate (9 weeks) and long-term (16 weeks) timepoints. Even though at least 25-fold fewer *HLF-ZsG* expressing cells were transplanted compared to reporter non-expressing or total *HLF-ZsG* targeted cells, comparable reconstitution levels (**Fig. 4c**) and similar lineage contributions (**Fig. 4d**) were achieved between these cohorts. Furthermore, we observed no difference between sgAAVS1 controls and *HLF-ZsG* targeted recipients, indicating that genetic manipulation of the *HLF* locus has no discernable functional impact on HSC activity aside from what can be attributed to RNP electroporation and rAAV6 transduction (**Fig. 4c first three panel columns**).

While these results already suggested a strong enrichment for reconstitution activity in reporter expressing cells, we additionally traced the engineered reporter allele in transplanted recipients by ddPCR to test whether human engraftment in the reporter non-expressing cohort had mainly emerged from non-targeted HSCs. This hypothesis was confirmed by the observation that HR allele ratios of 24.6% in the reporter non-expressing fraction at transplantation (**Fig. 4e, blue**) dropped sharply in the progeny of these cells *in vivo* (**Fig. 4f and g**). This was in contrast to total targeted or reporter expressing populations, where HR allele frequencies remained largely stable between pre- and post-transplantation timepoints. Strikingly, HR allele frequency reached 88.5% in the sorted *HLF-ZsG* expressing fraction indicating that most of these cells carried bi-allelic reporter integration (**Fig. 4e**).

Taken together, these results demonstrate that *HLF*-expression can be used to efficiently identify multipotent cells with high *in vivo* regenerative potential in $CD34^+$ cord blood cell cultures.

***HLF*-expression labels HSCs with extensive self-renewal capacity.**

At three weeks post-transplantation, a well-defined reporter positive sub-fraction within CD34-high bone marrow cells was detectable in recipients transplanted with total and ZsG-positive *HLF-ZsG* targeted cells but was entirely absent in ZsG-negative recipients (**Fig. 5a**). To test whether *HLF* expression continued to label stem cells after transplantation, we selected three *HLF-ZsG* targeted ZsG^+ primary recipients with detectable ZsG^+ cells at week 10.5 (**Fig. 5b**) and high (>96% each) HR allele ratios (**Fig. 4f**) as donors for secondary transplantation. As presented in **Fig. 5c**, we first magnetically enriched $CD34^+$ cells from the pooled bone marrow of these donors and then sorted reporter-expressing cells as well as a corresponding population that expressed similar CD34 levels but was 16.6-fold more abundant. Corresponding cell numbers of these populations (900 and 1.5*10e4, respectively) were intra-hepatically transplanted into newborn secondary recipient mice (**Fig. 5c**). Significant differences in human engraftment between HLF-ZsG^+ and $CD34^{high}$/HLF-ZsG^- secondary recipients were observed at short-term, intermediate and long-term post-transplantation timepoints (**Fig. 5d**). More specifically, while all HLF-ZsG^- secondary recipients were characterized by transient, low-level and exclusively myelo-monocytic reconstitution, pronounced chimerism with multilineage distribution was detectable in at least four of ten HLF-ZsG^+ recipients as long as 16 weeks post-transplantation (**Fig. 5d and e**).

Based on these observations, we conclude that in addition to labeling human HSCs *ex vivo*, *HLF* expression continues to mark haematopoietic stem cells with extensive reconstitution activity *in vivo*. These results therefore demonstrate the universal utility of *HLF* reporter transgenesis to visualize the human blood stem cell state in real-time under experimental conditions.

Discussion:

With the aim to develop an approach to directly mark the rare stem cell fraction within the human blood system, we identify *HLF* as the most selectively expressed gene in human HSCs, corroborating its reported roles in the transcriptional regulation of HSC self-renewal and multipotency in mice[36,37]. Interestingly, in addition to these endogenous activities, ectopically expressed HLF has the capacity to impart self-renewal to differentiation-committed blood cells through its DNA binding activity, either in the context of recurrent chromosomal fusions with *TCF3* in t(17;19) acute B-lymphoblastic leukemia[40] or as a reprogramming factor in murine induced haematopoietic stem cells[41], clearly positioning it as one of the most important master transcription factors in HSCs.

Highly specific transgenic labeling of mouse HSCs has recently overcome many of the challenges associated with accurately identifying and studying these cells in their physiological and anatomical contexts during development[13] and in the adult[10-12,14]. With this study, we provide the first demonstration of transgenic labeling of the human HSC state to date, paving the way for targeted genetic studies that, to our knowledge, have not been reported in human adult stem cells.

While the presented experimental procedure is currently not without its limits, most notably the functional impact on HSC vitality following nucleofection, we provide a directly quantifiable platform to optimize gene editing in HSCs under function-preserving conditions, an important remaining issue in the field.

Furthermore, quantitative readout of the human HSC reporter will likely find widespread utility in efforts to further optimize human HSC expansion conditions, either through screening for pharmacological self-renewal agonists or through systematic optimization of media composition and overall culture design.

It can be anticipated that the presented reporter strategy will serve as a powerful tool to systematically dissect and rebuild the human HSC state analogous to the approach that led to the discovery of induced pluripotency using an embryonic stem cell specific reporter[42].

Figure legends:

1) Specific Expression of HLF expression in enriched human HSC populations.

a) HLF expression is enriched in cultured human HSC subsets. Differentially expressed genes from CD201+ [7] and ITGA3/CD201+ [21] HSC-enriched population transcriptomes were intersected to identify consistently up or downregulated genes. Ranking based on fold-change between the most enriched (ITGA3$^+$/CD201$^+$) and most depleted (CD201-) HSC populations is summarized by a waterfall plot (left, log2-transformed). Range of expression is provided for each gene in square brackets (TPM = transcripts per kilobase million).

b) HLF-expression is undetectable in blood leukocyte populations. (curated from [22]).

c) Single cell transcriptomic overview of cord blood cell populations. Fresh and UM171-expanded CD34$^+$ (d7) (2 samples each condition, 15921 cells total), were scRNA sequenced (10X Chromium), integrated and clustered using Seurat 3[43]). Ten cell clusters were identified: haematopoietic stem/progenitor cells (HSPC), lymphoid-primed multipotent progenitors (LMPP), multipotent progenitors (MPP), granulo-monocytic progenitor (GMP), megakaryocyte-erythroid-mast cell progenitors (MEMP), megakaryocytes (mega), eosinophil/basophils (eo/ba), mast cells (mast), erythroid lineage cells, neutrophils (neutro) and monocytic/dendritic cells (mono/dc). Dimensional reduction was calculated using SPRING[44].

d-e) HSPC specific HLF expression in fresh CD34$^+$ cells and after 7d ex vivo expansion in the presence of UM171. *HLF* expression is shown in single cell transcriptomes split up by treatment. Normalized expression data (z-score, after MAGIC imputation[45]) is expressed in color scale.

f) Comparison of HLF expression specificity versus selected HSC-associated genes and common HSC surface marker genes. Gene-wise z-score distribution by treatment (fresh CD34+ and UM171 d7) is represented as density for each cell community (same color-code as in Fig. 1c). Mean z-score for HLF in HSPC cluster is provided as dotted reference line for each treatment.

g-i) HLF expression is strongly enriched in HSC clusters in human bone marrow.

g) Overview of cell clusters. (Human Cell Atlas; 101,935 cells; integrated data from eight donors, UMAP reduction; preprocessed data, clusters and labels adopted from[32]).

h) HLF expression. (z-score normalized, after MAGIC imputation).

i) Expression summary of selected HSC-associated genes in human bone marrow. (scaled expression averaged for each haematopoietic cell population and donor (n=8) in the dataset, row-normalized color scale).

j-l) HLF expression is restricted to HSC/MPP cluster in haematopoietic human fetal liver cells.

j) Overview of cell communities. (Human Cell Atlas; 113,063 cells; integrated data from 14 fetal livers across four developmental stages, UMAP reduction; preprocessed data, clusters and labels adopted from[33]). Mac, macrophage; Neut-my, neutrophil–myeloid; Mono-mac, monocyte-macrophage; Early L/TL, early lymphoid/T lymphocyte; pro., progenitor.

k) *HLF* expression (z-score normalized, after MAGIC imputation).

***l) Expression summary of HSC-associated genes* in fetal liver** (scaled expression was averaged for each haematopoietic cell population and four gestation stages (7-8, 9-11, 12-14 and 15-17 post-conception weeks), row-normalized color scale).

2) Engineering of a human genomic *HLF*-reporter.

a) Outline of the HLF-reporter targeting strategy using CRISPR/Cas9 and rAAV6. A site-specific DSB at the HLF stop codon (orange) located in exon 4 is generated by a Cas9/sgHLF ribonucleoprotein (RNP) complex. This stimulates homologous recombination (HR) with a single-

stranded donor template delivered through rAAV6 infection. The resulting HR event results in a transgenic locus that co-expresses the HLF open reading frame and a multifunctional ZsGreen (ZsG) expression cassette connected the endogenous HLF open reading frame by an EMCV internal ribosome entry site (ires). Grey boxes, HLF exons; white boxes, 5' and 3' untranslated regions; 3'/5' HA, homology arms; purple box, puromycin resistance or truncated EGFR (tEGFR) sequence linked to ZsG by a P2A for optional drug or antibody mediated selection; WPRE, Woodchuck Hepatitis Virus Post-transcriptional Response Element to improve transcript stability; pA, endogenous HLF polyadenylation signal.

b-e) Validation of the HLF reporter in human cell lines. HepG2 (HLF-expressing) and HEK293 (HLF non-expressing) cells were electroporated with Cas9/sgRNA RNP either as summarized in (a) or using sgAAVS1 as control. *HLF-ZP*, rAAV6 encoded HLF repair template driving expression of ZsG and Puromycin resistance.

b) Droplet digital PCR genotyping of targeted cell lines. Black dots represent HR-negative and red dots represent HR-positive PCR droplets. HR percentages (printed in red) were calculated as HR-positive divided by the total number of specific amplicon-containing droplets (black and red).

c) ddPCR strategy. ext. FW, external forward primer binding to a common region outside the 5' HA; 3' RV (reverse) primer amplifying unrecombined locus; ires RV (reverse) primer amplifying recombined locus; HR-negative and positive amplicons are detected by a common FAM-labeled probe and HR-positive amplicons are additionally recognized by a HEX-labeled probe that binds to the ires region of the transgene.

d) FACS analysis to detect reporter expression.

e) HLF expression levels in selected cell lines. Data curated from Human Protein Atlas [46].

f) Outline of experimental strategy to optimize reporter functionality in CD34$^+$ cord blood cells.

g) FACS sorting strategy to enrich/deplete HSCs based on CD201 expression from expanded CD34$^+$ cells at d3 of culture.

h) Selective HLF-reporter expression in HSC-containing subfractions of CD34+ cord blood cell cultures. HR allele frequencies were determined from one of four replicate wells. Total and ZsG$^+$ cell counts were determined by FACS on d7 and are normalized to 10e4 cells plated per 96-well post-electroporation at d4. MOI, multiplicity-of-infection.

i) Effect of TP53 knock-down on HR and cell survival. CD201$^+$ cells were sorted and targeted as in (h) (MOI400), and as an additional condition, electroporated with RNP and siRNA against TP53.

3) Selective HLF-reporter expression in human cord blood derived LT-HSC populations.

Cord blood derived CD34+ cells were processed as in **Fig. 2f** with the addition of siTP53 and transduction of rAAV6 *HLF-ZsG P2A tEGFR* at MOI400.

a) HR allele frequencies in CD201$^+$/$^-$ pre-sorted fractions as determined by ddPCR. Gated on FAM+ (common probe) droplets, HEX+ droplets (red) identify HR allele amplicons.

b-c) Reporter expression in CD201$^+$/$^-$ pre-sorted fractions. Aggregated FACS analysis (b) and summary by repeat (n=4 for CD201- and n=5 for CD201+ sorted, unpaired two-sided Wilcoxon test p-value is indicated) in (c).

d-e) Immuno-phenotypes of ex vivo expanded (+UM171) HLF-targeted HSPCs. FACS analysis of total (black) versus reporter expressing (green) populations at day 7. Percentages of increasingly restricted HSC gates are provided for each population. Aggregated FACS data in (d) and summary by repeat in (e).

f) Dimensional reduction based on FACS analysis. UMAP reduction using CD34, CD45RA, CD201, CD90, ITGA3 and ZsG FACS intensities from (d) was calculated and is represented as

2d density plot of all cells (grey, n = 306,797). Cells falling into HSC- or ZsG-gates are overlayed and color-coded as in (d).

g-h) HLF-reporter expression in immuno-phenotypic HSC gates. Reverse gating of the same data as above showing reporter expression in increasingly restricted HSC gates. Aggregated FACS data from all repeats in (g), summary in (h).

i) Extrapolated HLF expression within immuno-phenotypic HSC gates. Ranges of cell fractions expressing *HLF* are summarized for each sub-population under consideration of measured allelic HR ratios of 20.2% (a). Lower boundaries are based on a scenario where all ZsG+ cells carry bi-allelic reporter integration, upper values are based on entirely mono-allelic integration.

4) HLF-reporter labels repopulating cells in CD34+ cord blood cell cultures.

a) FACS plots showing the sorting of HLF-reporter targeted population for transplantation. *rAAV6 HLF-ZE*: recombinant rAAV6 particle encoding an *HLF* repair template with ires ZsGreen P2A tEGFR cassette.

b) Summary of transplantation layout. Transplantation cohorts and cell doses are represented using the same color-code as in (a).

c) Human engraftment summary of transplanted NSGS recipients. Human bone marrow chimerism determined based on human CD45+ cells among total CD45+ (mouse and human) cells at short (week 3), intermediate (week 9) and long-term (week 16) post-transplantation timepoints is plotted using the same color code as in (a) and (b). Each recipient mouse is represented along the x-axis (NSGS-ID). Recipients are arranged by descending average reconstitution across all timepoints. Recipients #25912, #25914 and #25916 were sacrificed at week 10 post-transplantation to be used as donors for secondary transplantation (summarized in Fig. 5).

d) Lineage proportion of transplanted recipients. Bone marrow biopsies were analyzed and are arranged along timepoints and the individual recipients as in (e). Normalized proportions of B-cells (CD19), myeloid cells (CD33) and T-cells (CD3) within human $CD45^+$ cells for each timepoint and recipient are color-coded as indicated.

e) HR allele frequencies in pre-transplanted cell populations. top panel: ddPCR droplets are pre-gated based on FAM-positivity, black droplets represent FAM+/HEX- events indicative of untargeted alleles, red droplets (FAM/HEX double positive) indicate targeted alleles, sub-sampled to 300 droplets per specimen. bottom panel, quantification summary of HR frequencies calculated based on targeted/(untargeted+targeted) droplets.

f) ddPCR analysis of bone marrow biopsies at weeks 3, 9 and 16. Specimens are arranged as in (c), ddPCR droplets are represented as in (e), sub-sampled to 50 droplets per specimen and timepoint.

g) HR allele tracing summary. Summarized data representation of (e) and (f). Dashed red lines represent allele frequencies at time of transplantation. Bars represent average HR allele frequencies from (e) with standard error bars, color-codes as in (a-c).

5) HLF-reporter labels human HSCs with extensive self-renewal capacity.

a) FACS plots of CD34+/HLF-ZsG+ population. Representative bone marrow biopsies of reporter-negative (*sgHLF/ rAAV6 HLF-ZE* targeted, ZsG^- sorted, left) and reporter-positive (*sgHLF/ rAAV6 HLF-ZE* targeted, ZsG^+ sorted, right panel) primary recipients, gated on human CD45+.

b) Summary of CD34+/HLF-ZsG+ population. Population overview of all primary recipients, pre-gated on human CD45+, recipient mice are arranged according to engraftment levels as in Fig. 4c.

c) Strategy for secondary transplantation. Bone marrow of three primary recipients (*sgHLF/ rAAV6 HLF-ZE* targeted, ZsG^+ sorted cohort) was pooled and magnetically enriched for human

CD34 expression. Reporter-expressing (ZsG+) and non-expressing cells (ZsG-) with comparable levels of CD34 expression were sorted for transplantation. Intra-hepatic transplantation into newborn NSGS recipients as outlined. Corresponding cell doses of HLF-ZsG+ (n=10) and HLF-ZsG- (n=7) were transplanted.

d) Human engraftment summary of secondary recipients. Human bone chimerism in indicated tissues was determined based on human CD45-expressing cells among total (mouse and human) CD45+ cells at short (week 5, blood), intermediate (week 9, marrow) and long-term (week 16, marrow and spleen) post-transplantation. Dashed line represents the 0.1% mark used as cut-off for engraftment positivity. Significance was calculated by unpaired, one-sided (alternative = “greater”) Wilcoxon test and is provided as p-value for a given comparison.

d) Lineage distribution of engrafted human cells. Positive specimens from (d) are shown and color-coded for B-cells (CD19+), myeloid cells (CD33+) and T-cells (CD3). Normalized for lineage proportions within human CD45+ cells. Samples with less than 0.1% of human chimerism are designated negative (neg.).

Extended Data legends:

E1) Intersection of EPCR and ITGA3 transcriptome datasets.

Differentially expressed genes in the EPCR (left volcano, Fares et al., Blood 2017) and ITGA3 dataset (right volcano, Tomellini et al., Cell Reports 2019) were intersected to determine commonly down-regulated (blue, n=7, listed below) and up-regulated (red, n=17, listed below) genes. p- and beta-values (Kallisto-Sleuth pipeline) for EPCR and ITGA3 were used to set cut-offs for differentially expressed genes in the respective datasets.

E2) Expression of HSC-associated genes and surface markers in fresh and UM171-expanded CD34+ cord blood cells.

Relative gene expression is shown in single cell transcriptomes split up by treatment. Normalized expression (z-score, after MAGIC imputation[45]) is expressed in color scale. Cell clusters are mapped in main Figure 1c.

E3) Single cell expression values of HSC- and lineage-associated genes in fresh and UM171-expanded CD34+ cord blood cells (Scaled).

Scaled expression (after SCT transformation) including cell meta data in tab-separated format.

E4) Single cell expression values of HSC- and lineage-associated genes in fresh and UM171-expanded CD34+ cord blood cells (Imputed).

Imputed expression (after MAGIC imputation, t=1 using all genes) including cell meta data in tab-separated format.

E5) Expression of HSC-associated genes and surface markers in human bone marrow and fetal liver.

Relative gene expression is shown in single cell transcriptomes in human bone marrow **(a)** and fetal liver **(b)**. Normalized expression (z-score, after MAGIC imputation[45]) is expressed in color scale. Respective cell clusters are mapped in main Figures 1 g (marrow) and j (fetal liver).

E6) High-efficiency of Cas9/sgHLF RNP indel formation at the target site.

CD34+ cord blood cells were pre-expanded in UM171-containing media and nucleofected with a Cas9/sgHLF RNP complex (Lonza 4d, program DZ100) after 3d of culture. After additional 3d of culture, the genomic region flanking the sgRNA target site was PCR amplified and sequenced by Sanger sequencing. Untreated cells were used as control. Analysis was done using the Synthego ICE Analysis webtool (https://ice.synthego.com/#/).

a) Experimental layout.

b) Efficient Indel formation. Sanger sequence chromatograms of control and Cas9/sgHLF RNP treated CD34+ cord blood cultures are shown. Indel frequency reached >90% based on ICE analysis.

E7) Gating strategies to analyze human reconstitution in transplanted host.

Numbers indicate consecutive gating steps.

E8) Gating strategies of secondary sort.

Numbers indicate consecutive gating steps. Gating step 4 was used to exclude auto-fluorescent cells.

Methods:

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Analyses of bulk transcriptomes. EPCR and ITGA3 specific datasets were previously reported. In brief, differentially expressed genes (DEGs) from the ITGA3 dataset were exactly as in Tomellini et al. (Cell Reports, 2019) using the Kallisto/Sleuth pipeline and the full GRCh38 v92 annotation (including non-coding genes). The EPCR dataset (Fares et al., Blood 2017) was re-analyzed in the same fashion. Expression weighted fold-change (beta-value) and p-value (sleuth) cut-off values were designated based on ITGA3 (beta ± 0.844; p = 0.0071) and EPCR (beta ±

1.412; p = 3.32e-7) in the respective datasets. Intersection of positive and negative DEGs yielded 17 and 7 genes respectively. Analyses and heatmaps were generated in R.

Analyses of single cell transcriptomes from fresh and UM171 expanded cord blood. CD34+ cord blood (CB) cells, either freshly thawed or culture-expanded with UM171 (35nM) were single-cell sequenced on a Chromium Single-Cell Controller (10X Genomics) using the Single Cell 3' Reagent Kit version 2 according to manufacturer's instructions. Target cell numbers were 6,000 per condition. scRNAseq libraries were sequenced on an Illumina NovaSeq device using a S2 (PE 28x91) setup or on a Illumina HiSEQ 4000 using 26x98 cycles. A standard Cellranger v3.0.1 pipeline was used for read mapping (GRCh38 annotation) and demultiplexing. Subsequent analyses were done in Seurat (v3) based on Cellranger prefiltered barcode/feature matrices and included (i) exclusion of cells with less genes or UMIs than the respective medians minus 2 standard deviations (ii) exclusion of cells with more genes or UMIs than the respective medians plus 2 standard deviations (multiplets), (iii) exclusion of cells with more than 7% mitochondrial gene expression (apoptotic cells). Expression counts were normalized using the SCTransform wrapper in Seurat including regression on cell cycle scores and mitochondrial gene content. Seurat integration was performed using the top 241 integration anchors (250 minus sex specific genes) in the first 30 dimensions, followed by PCA dimensional reduction, FindNeighbors and FindClusters (resolution = 0.5) in the first 15 integrated dimensions. SPRING embedding was calculated on the integrated expression matrix using the SPRING webtool (https://kleintools.hms.harvard.edu/tools/spring.html). For visualization, data imputation was calculated on SCT transformed data of all genes using the MAGIC wrapper (t = 1) in Seurat.

Analysis of Human Cell Atlas single cell transcriptomes. For the human bone marrow dataset, count matrices were directly downloaded through the 'HCAData' package in R using HCAData('ica_bone_marrow'). Preprocessing and cell annotations were adopted based on

consolidated grouping metadata from Hay et al. 2018. In Seurat, the full data set was split up by donor id (MontonBM id) and sequencing run, then SCT transformed including regression of mitochondrial content before sample integration. The pre-processed human fetal liver dataset was kindly provided as annotated count matrix by Muzlifah Haniffa (Newcastle, UK). Further processing in Seurat was done as with the bone marrow dataset, except split up by project id. Both datasets were integrated using 2000 anchors in the first 50 dimensions and the reciprocal PCA workflow. Dimensional reduction was calculated using PCA and UMAP of the first 50 dimensions. Cluster annotations and labels were adopted as published for both datasets. For visualization purposes, data imputation was calculated on SCT transformed data of all genes using the MAGIC wrapper (t = 1) in Seurat. Heatmap representation is based on average SCT transformed expression per cluster and donor in bone marrow, or per cluster and gestation stage in fetal liver dataset.

AAV vector cloning and production. A 2882bp XbaI/NotI backbone fragment of the AAV2 SaCas9/sgRNA expression vector pX601 (Addgene #61591) was used to clone a synthetic DNA fragment (IDT) with 5' and 3' homology arms (485 and 496 bp around HLF stop codon), flanking a small polylinker region (AscI and BglII restriction sites). Gibson assembly was used to insert different reporter cassettes directly downstream of the HLF stop codon using the AscI and BglII linearized vector. Recombinant AAV2 pseudotype 6 vector production was done by the Viral Vector Facility of the Canadian Neurophotonics Platform (Laval University, Quebec, Canada) and routinely reached titers of more than 3*10e12 vector genomes per ml as determined by Q-PCR.

Human CD34+ cord blood cells. Use of human cord blood specimens was approved by the Research Ethics Boards of the University of Montreal, Maisonneuve-Rosemont Hospital (Montreal, QC, Canada) and Charles LeMoyne Hospital (Greenfield Park, QC, Canada). Umbilical cord blood units were collected from consenting mothers according to an ethically approved

protocol at Charles-LeMoyne Hospital (Greenfield Park, QC, Canada). Human CD34+ cord blood (CB) cells were isolated using the EasySep Human Cord Blood CD34 Positive Selection Kit II (StemCell Technologies #17896). Initially, cord blood volume was measured and 5 mL of RosettaSep cocktail was added for each ml of sample and incubated for 20 min at RT. The CB sample was then diluted 1:1 with recommended medium (PBS, 2mM EDTA, 2% FBS, 1ml/ml DNase). A maximum 30 ml of diluted cord blood was carefully layered on 15 mL of Lymphoprep (StemCell Technologies #07851) in a 50ml falcon tube for density gradient centrifugation (StemCell Technologies #85450) and centrifuged for 20 minutes, 1200 g, without brakes. The middle layer containing mononuclear cells was harvested and transferred to a new 50 ml tube, topped up with the recommended medium and centrifuged for 10 minutes, 300 g, brakes low. Cell pellets were collected and resuspended in the recommended medium according to the starting volume of the CB sample. Immunomagnetic labeling and separation were performed according to the manufacturer's manual of the EasySep Human Cord Blood CD34 Positive Selection Kit II (StemCell Technologies #17896). Cells were either used immediately for experiments or frozen in FBS 10% DMSO.

CD34+ enriched cord blood cell culture. Culture of isolated human CB-derived CD34+ cells was performed in HSPC expansion media comprised of Stem Span serum-free media (StemCell Technologies #09855) supplemented with Glutamax (Invitrogen #35050061), low density lipoprotein (StemCell Technologies #02698), 100 ng/ ml SCF (Shenandoah Biotechnology #100-04), 100 ng/ml Flt3L (Shenandoah Biotechnology #100-21), 50 ng/ml TPO (R&D system #288-TP) and 35 nM UM171 (StemCell Technologies #72914).

Nucleofection and transduction of cell lines. The chemically modified synthetic guide RNA targeting HLF (sgHLF: 5'-AGGCACGGGCCCCUGUAGGA-3'; PAM: TGG) was custom purchased from Synthego or Trilink. The control sgRNA against AAVS1 (sgAAVS1: 5'-

GGGGCCACUAGGGACAGGAU-3'; PAM: TGG) was purchased from Synthego. Recombinant Alt-R HiFi S.p. Cas9 Nuclease V3 (IDT, #1081059) was used throughout. All nucleofections were done using a 4D Nucleofector Core and X Unit (Lonza #AAF-1002B and #AAF-1002X) in 1M buffer (5 mM KCl, 15 mM $MgCl_2$, 120 mM Na_2HPO_4/NaH_2PO_4, pH 7.2, 50 mM mannitol) or P3 buffer in either 100 ul cuvettes (Lonza #V4XP-3024) or 2x8 20 ul cuvette strips (Lonza #V4XP-3032). Nucleofection of HepG2 (ATCC #HB-8065) and HEK293T (Clontech #632617) was done in 100 ul cuvettes and 1M buffer, with 3ug Cas9 and 8ug of sgRNA and 10e6 cells, using program EH100 for HepG2 and program CM130 for HEK293T cells. After nucleofection, the cells were immediately plated in media containing 45,000 MOIs of reporter encoding rAAV6. Cells were then expanded and frozen, FACS and ddPCR data was acquired two weeks post-nucleofection.

Nucleofection and transduction of CD34+ cells. Nucleofection of was carried out with 3ug Cas9 and 8ug of sgRNA and 10e6 cells per 100 ul, using the DZ100 program. 4 days before nucleofection, CD34+ cells were thawed and plated at 1.5*10e5 cells/ml in HSPC expansion media containing UM171 (35 nM). At day 3, cells had typically expanded 2-3fold, and were stained with anti-CD34-BV421 (BD #562577, 1:50) and anti-CD201-APC (Biolegend #351906, 1:100) and sorted based on CD201 expression. The sorted cells were plated back in culture at 1.5*10e5 cells/ml. 24h later, cells were harvested, washed in PBS and taken up in 1M nucleofection buffer containing 11ug the pre-assembled Cas9 sgRNA RNP complex and were indicated 20 fmol of siRNA against TP53 (ThermoFisher #4390825, siRNA id s605). After nucleofection, the cells were immediately plated at 2-4*10e5 cell per ml in HSPC media optionally containing 400 MOIs of reporter encoding rAAV6. Half media changes were done on days 5 and 6, analysis or transplantation was done on day 7.

Assessment of homologous recombination using droplet digital PCR (ddPCR). Primers were designed to amplify a 605bp region of the wildtype or 601bp region of the targeted allele

using a set of 3 primers. A common external forward primer (ext_FW; 5'-CCACCTGCTTTCATCCAGC-3') binds in intron 3 upstream of the left homology arm, while the reverse primer binds either in the right homology arm (3'_RV; 5'-GGTAAAGTGCTGATGTCAGAAAGG-3') or in the IRES region of the reporter (ires_RV; 5'-TAACATATAGACAAACGCACACCG-3'). The 3'_RV primer binds 77bp downstream of the Cas9 cut site, thus amplifying most non-integrated alleles. A FAM labeled non-fluorescent quencher dual-labeled probe (HLF_common_FAM_probe; 5'-FAM-TCTGATCTCTGCTTCACTGAGCACGC-ZEN-Iowa-black-3') binds the HLF locus within the left homology arm and as such will detect both the wildtype and targeted allele amplicons. A second probe labelled with HEX and non-fluorescent quencher (ires_HEX_probe; 5'-HEX-CAAGCGGCTTCGGCCAGTAACGTTAG-ZEN-Iowa-black-3') binds in the IRES cassette and will detect only targeted allele amplicons. PCR efficiency and specificity were validated in standard PCR assays prior to ddPCR using synthetic gBlock DNA fragments consisting of synthetic wildtype or targeted alleles, combined with primer drop-out testing. Oligonucleotides and gBlocks were synthesized by IDT. To assess targeting of the HLF locus following electroporation and in xenograft recipients, approximately 30,000 cells were lysed in 25ul Quick Extract Buffer (Lucigen #QE09050), according to the manufacturer's directions. Briefly, cells were pelleted and resuspended in lysis buffer, then heated at 65° for 6min, vortexed and incubated for 2min at 95°. For cell line experiments with a larger cell numbers, the amount of lysis buffer was scaled up accordingly. 22ul ddPCR reactions contained 2.2ul of cell lysate, 11ul of 2X ddPCR Supermix for Probes (No dUTP, BioRad #1863024), 1.1ul (4U) FastDigest-HindIII (prediluted 1/3) (Thermo Fisher) and 11.1ul 20X primer-probe mix. Final concentration of primers and probes were 0.25uM and 0.9uM each, respectively. ddPCR reaction emulsion was created in a DG8 cartridge with DG8 gasket (BioRad #1863009) from 20ul PCR mix and 70ul droplet generation oil for probes (BioRad #1863052), using a QX200 Droplet Generator (BioRad #1864002). 40ul of PCR emulsion was pipetted into a 96-well QX200 compatible PCR plate (BioRad #12001925), which was covered

with a Pierceable Foil Heat Seal and sealed in a PX1 PCR plate sealer (BioRad #1814000) before cycling in a C1000 deep-well thermocycler (BioRad #1851197). Thermocycler conditions were as follows: 95°C for 10min; 50 cycles of 30sec at 95°C, 2min at 58°C, 2min at 72°C, followed by 10 min at 98°C, with a ramp speed of 2°C / step throughout. ddPCR products were measured using a QX200 Droplet reader (BioRad #1864001). Amplitude and Cluster data was exported from QuantaSoft Software (BioRad #1864003) and analyzed in R using the ddPCR package.

Flow cytometry. Fresh or cultured cells were analyzed using a FACS CantoII (BD Biosciences) and sorted on a BD FACSAria II cell sorter (BD Bioscience). For phenotypic HSPC analysis, cells were stained using the following antibodies: anti-CD34 (BV421, BD Biosciences #562577, 1:50), anti-CD45RA (APC-Cy7, BioLegend #304128, 1:50), anti-EPCR (PE, BioLegend #351904, 1:100), anti-CD90 (PE-Cy7, BioLegend #328124, 1:150), anti-ITGA3 (APC, eBioscience #17-0494-42, 1:50). Flow cytometry data was analyzed by FlowJo v10 software. UMAP projection was calculated in FlowJo using the UMAP plugin.

Transplantation of targeted HSPCs and analysis of engraftment. CB cells were transplanted by tail vein injection into sublethally irradiated (250 cGy) 8-16 week-old female NOD-Scid IL-2Rɣ null 3/GM/SF (NSGS, The Jackson Laboratory) mice. The number of cord blood derived cells transplanted per mouse as well as detailed engraftment data for each mouse is provided in figures 4 and 5. The engraftment of human cells in the bone marrow of NSGS mice was monitored by flow cytometry at the indicated post-transplantation timepoints. Bone marrow (BM) cells were collected by femoral aspiration (first two timepoints per experiment) or by flushing the 2 femura and tibiae at experimental timepoints. Flow cytometry analysis was performed on freshly collected BM cells. Cells were treated with red blood cell lysis buffer (StemCell Technologies #20110), washed and stained with the following antibody cocktails: 3 weeks timepoint in primary recipients: anti-mouse-CD45.1 (APC-eFluor780, eBioscience #47-0453-82), anti-human-CD45 (Pacific

Blue, BioLegend #304029), anti-CD33 (PE, BD Biosciences #555450), anti-CD19 (PE-Cy7, BD Biosciences #557835), anti-CD34 (APC, BD Biosciences #555824); 9 week timepoint in primary recipients: anti-human-CD45 (Pacific Blue, BioLegend #304029), anti-CD33 (PE, BD Biosciences #555450), anti-CD19 (PE-Cy7, BD Biosciences #557835), anti-CD34 (APC, BD Biosciences #555824), anti-CD3 (APC-Cy7, BD Biosciences #557832); 16 week timepoint in primary recipients and all timepoints in secondary recipients: anti-mouse-CD45.1 (APC, BioLegend #110714), anti-human-CD45 (Pacific Blue, BioLegend #304029), anti-CD33 (PE, BD Biosciences #555450), anti-CD19 (PE-Cy7, BD Biosciences #557835), anti-CD34 (APC, BD Biosciences #555824), anti-CD3 (APC-Cy7, BD Biosciences #557832). For secondary transplantations, the bone marrow from three primary recipients was harvested (hips, femura, tibiae), pre-enriched for human CD34 expression using the EasySep Human Cord Blood CD34 Positive Selection Kit II (StemCell Technologies #17896), then sorted as outlined and transplanted into the livers of one-day old NSGS recipients directly after 100 cGy irradiation.

Data Availability. The accession numbers for the previously published datasets are as follows: EPCR population dataset (GSE77128), ITGA3 population dataset (GSE130974), the human bone marrow single cell dataset is available through the Human Cell Atlas consortium and was downloaded as count matrix through the HCAData portal in R (https://github.com/federicomarini/HCAData). The human fetal liver dataset was obtained as annotated count matrix from Muzlifah Haniffa or else is available at ArrayExpress with accession code E-MTAB-7407. Fresh and UM171-expanded CD34+ single-cell datasets have been deposited to GEO under accession GSE153370.

Acknowledgments. We thank Keith Humphries, Julie Lessard and Trang Hoang for critical reading of the manuscript, Annie Gosselin and Angelique Bellemare for assistance in FACS

sorting, Melanie Frechette and Valérie Blouin-Chagnon for assistance with mouse experiments, and Mike Tyers for sharing nucleofection equipment.

Author contributions. B.L.: project conception, designed and performed experiments, experimental and bioinformatic data analyses and interpretation, generated all figures, wrote manuscript; T.M. designed and performed ddPCR analyses; J.C. and E.T.: technical setup and assistance with CD34+ cell culture and FACS experiments, interpretation of results; S.C., I.B.: assistance with CD34+ cell purification and banking; N.M.: assistance with CD34+ cell purification, banking and mouse experiments; G.S.: project supervision and coordination, experimental design, interpretation of results and manuscript preparation.

Competing interests. The authors declare no competing interests.

Citations:

1. Till, J. E. & McCulloch, E. A. A Direct Measurement of the Radiation Sensitivity of Normal Mouse Bone Marrow Cells. *Radiat. Res.* **14,** 213 (1961).
2. Siminovitch, L., McCulloch, E. A. & Till, J. E. The distribution of colony-forming cells among spleen colonies. *J Cell Comp Physiol* **62,** 327–336 (1963).
3. Wu, A. M., Till, J. E., Siminovitch, L. & McCulloch, E. A. A cytological study of the capacity for differentiation of normal hemopoietic colony-forming cells. *Journal of Cellular Physiology* **69,** 177–184 (1967).
4. Wilkinson, A. C., Igarashi, K. J. & Nakauchi, H. Haematopoietic stem cell self-renewal in vivo and ex vivo. *Nat Rev Genet* **132,** 1–14 (2020).
5. Doulatov, S., Notta, F., Laurenti, E. & Dick, J. E. Hematopoiesis: A Human Perspective. *Cell Stem Cell* **10,** 120–136 (2012).

6. Dorrell, C., Gan, O. I., Pereira, D. S., Hawley, R. G. & Dick, J. E. Expansion of human cord blood CD34+CD38−cells in ex vivo culture during retroviral transduction without a corresponding increase in SCID repopulating cell (SRC) frequency: dissociation of SRC phenotype and function. *Blood* **95,** 102–110 (2000).

7. Fares, I. *et al.* EPCR expression marks UM171-expanded CD34+ cord blood stem cells. *Blood* **129,** 3344–3351 (2017).

8. Frelin, C. *et al.* GATA-3 regulates the self-renewal of long-term hematopoietic stem cells. *Nat Immunol* **14,** 1037–1044 (2013).

9. Gazit, R. *et al.* Fgd5 identifies hematopoietic stem cells in the murine bone marrow. *J Exp Med* **211,** 1315–1331 (2014).

10. Acar, M. *et al.* Deep imaging of bone marrow shows non-dividing stem cells are mainly perisinusoidal. *Nature* **526,** 126–130 (2015).

11. Chen, J. Y. *et al.* Hoxb5 marks long-term haematopoietic stem cells and reveals a homogenous perivascular niche. *Nature* **530,** 223–227 (2016).

12. Xiang, P. *et al.* A knock-in mouse strain facilitates dynamic tracking and enrichment of MEIS1. *Blood Advances* **1,** 2225–2235 (2017).

13. Yokomizo, T. *et al.* Hlf marks the developmental pathway for hematopoietic stem cells but not for erythro-myeloid progenitors. *J Exp Med* **280,** jem.20181399–320 (2019).

14. Christodoulou, C. *et al.* Live-animal imaging of native haematopoietic stem and progenitor cells. *Nature* **514,** 1–6 (2020).

15. CRISPR/Cas9 β-globin gene targeting in human haematopoietic stem cells. *Nature* **539,** 384–389 (2016).

16. Boitano, A. E. *et al.* Aryl hydrocarbon receptor antagonists promote the expansion of human hematopoietic stem cells. *Science* **329,** 1345–1348 (2010).

17. Fares, I. *et al.* Pyrimidoindole derivatives are agonists of human hematopoietic stem cell self-renewal. *Science* **345,** 1509–1512 (2014).

18. Chaurasia, P., Gajzer, D. C., Schaniel, C., D'Souza, S. & Hoffman, R. Epigenetic reprogramming induces the expansion of cord blood stem cells. *J Clin Invest* **124,** 2378–2395 (2014).

19. Mantel, C. R. *et al.* Enhancing Hematopoietic Stem Cell Transplantation Efficacy by Mitigating Oxygen Shock. *Cell* **161,** 1553–1565 (2015).

20. Bai, T. *et al.* Expansion of primitive human hematopoietic stem cells by culture in a zwitterionic hydrogel. *Nat Med* **25,** 1566–1575 (2019).

21. Tomellini, E. *et al.* Integrin-α3 Is a Functional Marker of Ex Vivo Expanded Human Long-Term Hematopoietic Stem Cells. *Cell Reports* **28,** 1063–1073.e5 (2019).

22. Uhlen, M. *et al.* A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* **366,** eaax9198 (2019).

23. Velten, L. *et al.* Human haematopoietic stem cell lineage commitment is a continuous process. *Nat Cell Biol* **19,** 271–281 (2017).

24. Haas, S., Trumpp, A. & Milsom, M. D. Causes and Consequences of Hematopoietic Stem Cell Heterogeneity. *Cell Stem Cell* **22,** 627–638 (2018).

25. Zheng, S., Papalexi, E., Butler, A., Stephenson, W. & Satija, R. Molecular transitions in early progenitors during human cord blood hematopoiesis. *Molecular Systems Biology* **14,** e8041 (2018).

26. Calvanese, V. *et al.* MLLT3 governs human haematopoietic stem-cell self-renewal and engraftment. *Nature* **576,** 1–6 (2019).

27. Aguilo, F. *et al.* Prdm16 is a physiologic regulator of hematopoietic stem cells. *Blood* **117,** 5057–5066 (2011).

28. Civin, C. I. *et al.* Antigenic analysis of hematopoiesis. III. A hematopoietic progenitor cell surface antigen defined by a monoclonal antibody raised against KG-1a cells. *J Immunol* **133,** 157–165 (1984).

29. Yin, A. H. *et al.* AC133, a Novel Marker for Human Hematopoietic Stem and Progenitor Cells. *Blood* **90,** 5002–5012 (1997).

30. Baum, C. M., Weissman, I. L., Tsukamoto, A. S., Buckle, A. M. & Peault, B. Isolation of a candidate human hematopoietic stem-cell population. *Proc Natl Acad Sci USA* **89,** 2804–2808 (1992).

31. Notta, F. *et al.* Isolation of Single Human Hematopoietic Stem Cells Capable of Long-Term Multilineage Engraftment. *Science* **333,** 218–221 (2011).

32. Hay, S. B., Ferchen, K., Chetal, K., Grimes, H. L. & Salomonis, N. The Human Cell Atlas bone marrow single-cell interactive web portal. *Exp Hematol* **68,** 51–61 (2018).

33. Popescu, D.-M. *et al.* Decoding human fetal liver haematopoiesis. *Nature* **574,** 365–371 (2019).

34. Bak, R. O., Dever, D. P. & Porteus, M. H. CRISPR/Cas9 genome editing in human hematopoietic stem cells. *Nature protocols* **13,** 358–376 (2018).

35. Wang, X. *et al.* A transgene-encoded cell surface polypeptide for selection, in vivo tracking, and ablation of engineered cells. *Blood* **118,** 1255–1263 (2011).

36. Komorowska, K. *et al.* Hepatic Leukemia Factor Maintains Quiescence of Hematopoietic Stem Cells and Protects the Stem Cell Pool during Regeneration. *Cell Reports* **21,** 3514–3523 (2017).

37. Wahlestedt, M. *et al.* Critical Modulation of Hematopoietic Lineage Fate by Hepatic Leukemia Factor. *Cell Reports* **21,** 2251–2263 (2017).

38. Garg, S. *et al.* Hepatic leukemia factor is a novel leukemic stem cell regulator in DNMT3A, NPM1, and FLT3-ITD triple-mutated AML. *Blood* **134,** 263–276 (2019).

39. Schiroli, G. *et al.* Precise Gene Editing Preserves Hematopoietic Stem Cell Function following Transient p53-Mediated DNA Damage Response. *Cell Stem Cell* **0,** (2019).

40. Inaba, T. *et al.* Fusion of the leucine zipper gene HLF to the E2A gene in human acute B-lineage leukemia. *Science* **257,** 531–534 (1992).

41. Riddell, J. *et al.* Reprogramming committed murine blood cells to induced hematopoietic stem cells with defined factors. *Cell* **157,** 549–564 (2014).

42. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126,** 663–676 (2006).

43. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **0,** 1888–1902.e21 (2019).

44. Weinreb, C., Wolock, S. & Klein, A. M. SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* **34,** 1246–1248 (2018).

45. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174,** 716–729.e27 (2018).

46. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356,** eaal3321 (2017).

Fig. 1: Specific Expression of *HLF* in enriched human HSC populations.

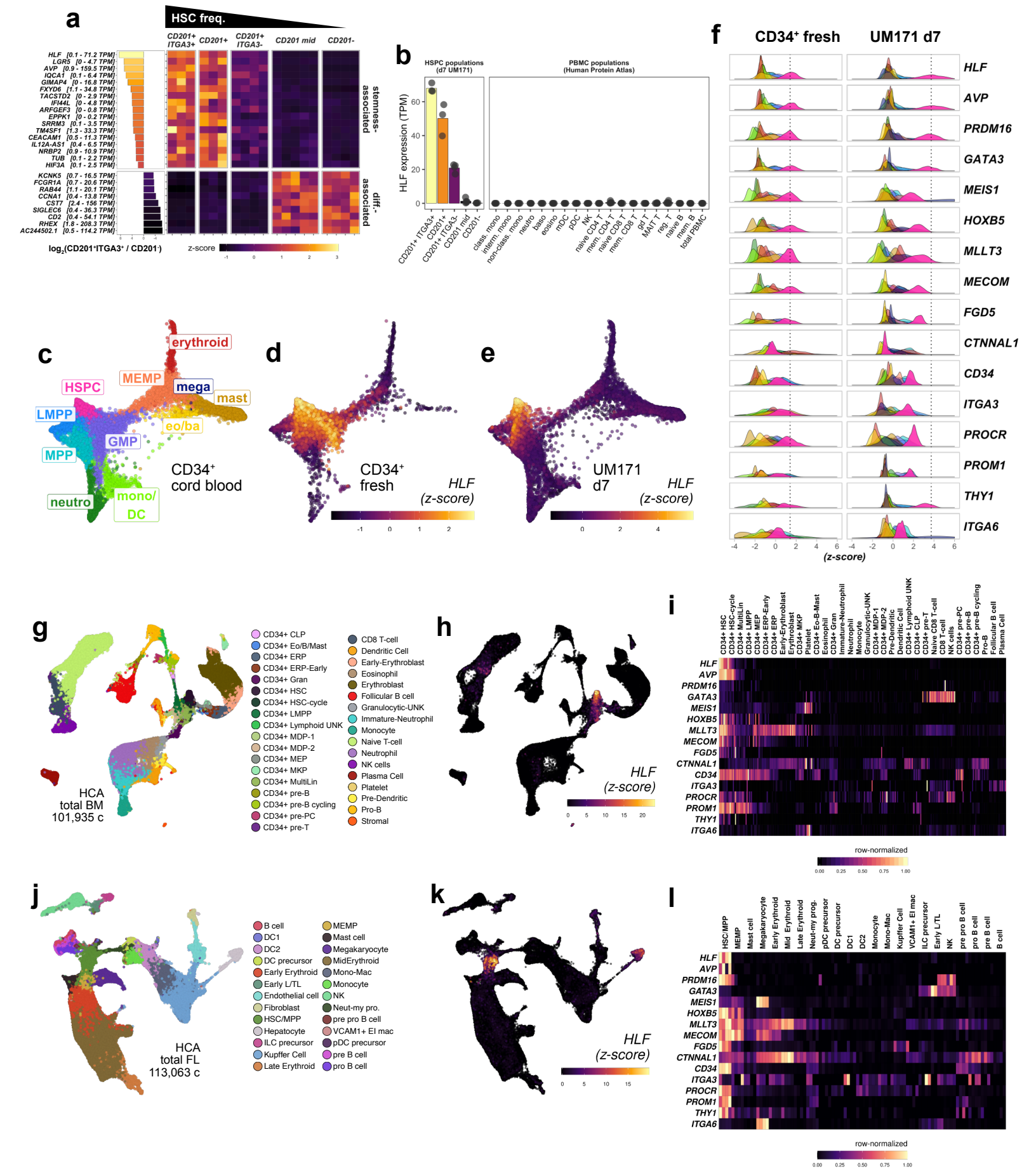

Fig. 2: Engineering of a genomic *HLF*-reporter transgene in human cells.

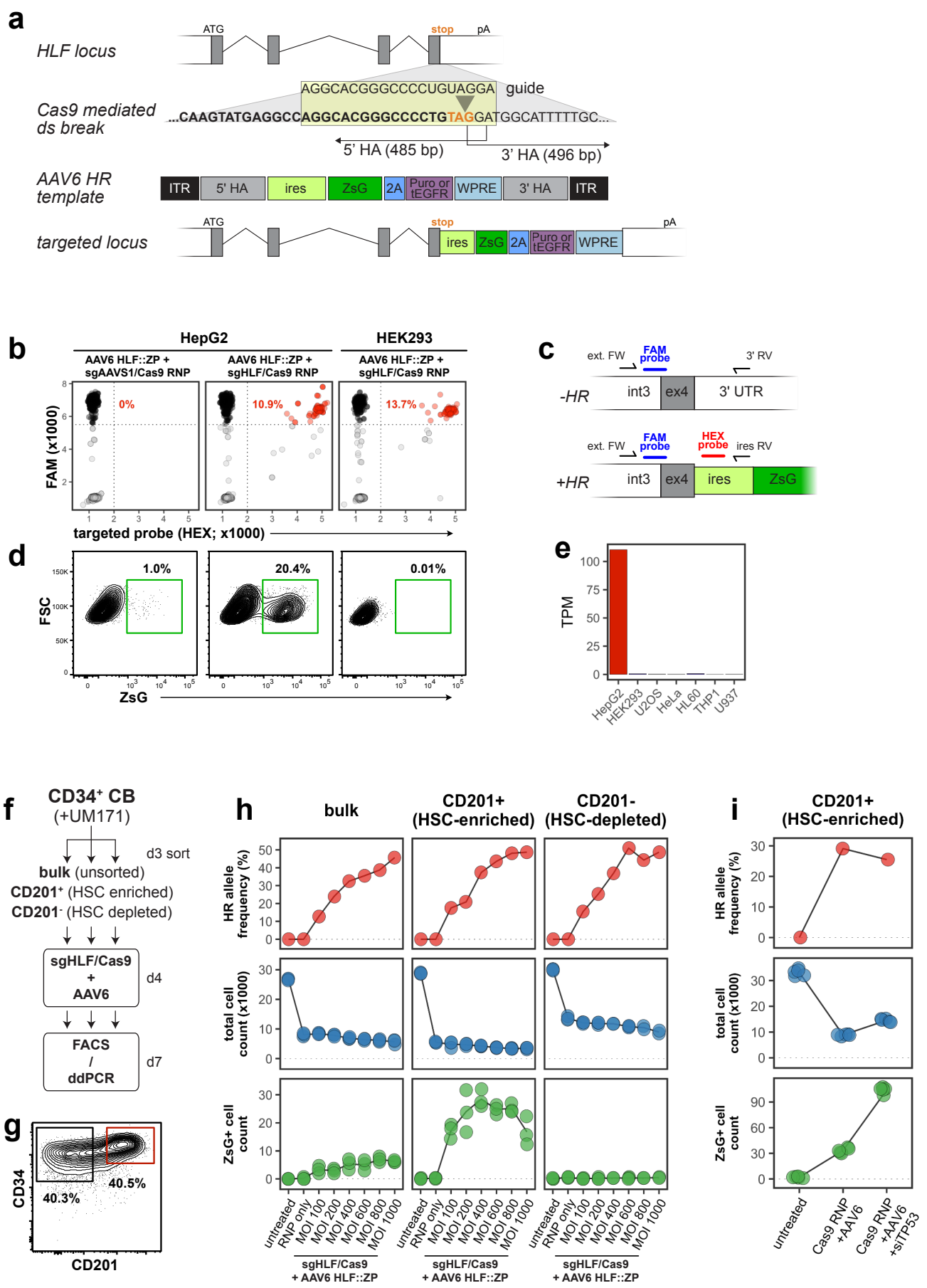

Fig. 3: Selective *HLF*-transgene expression in immunophenotypic human LT-HSC populations.

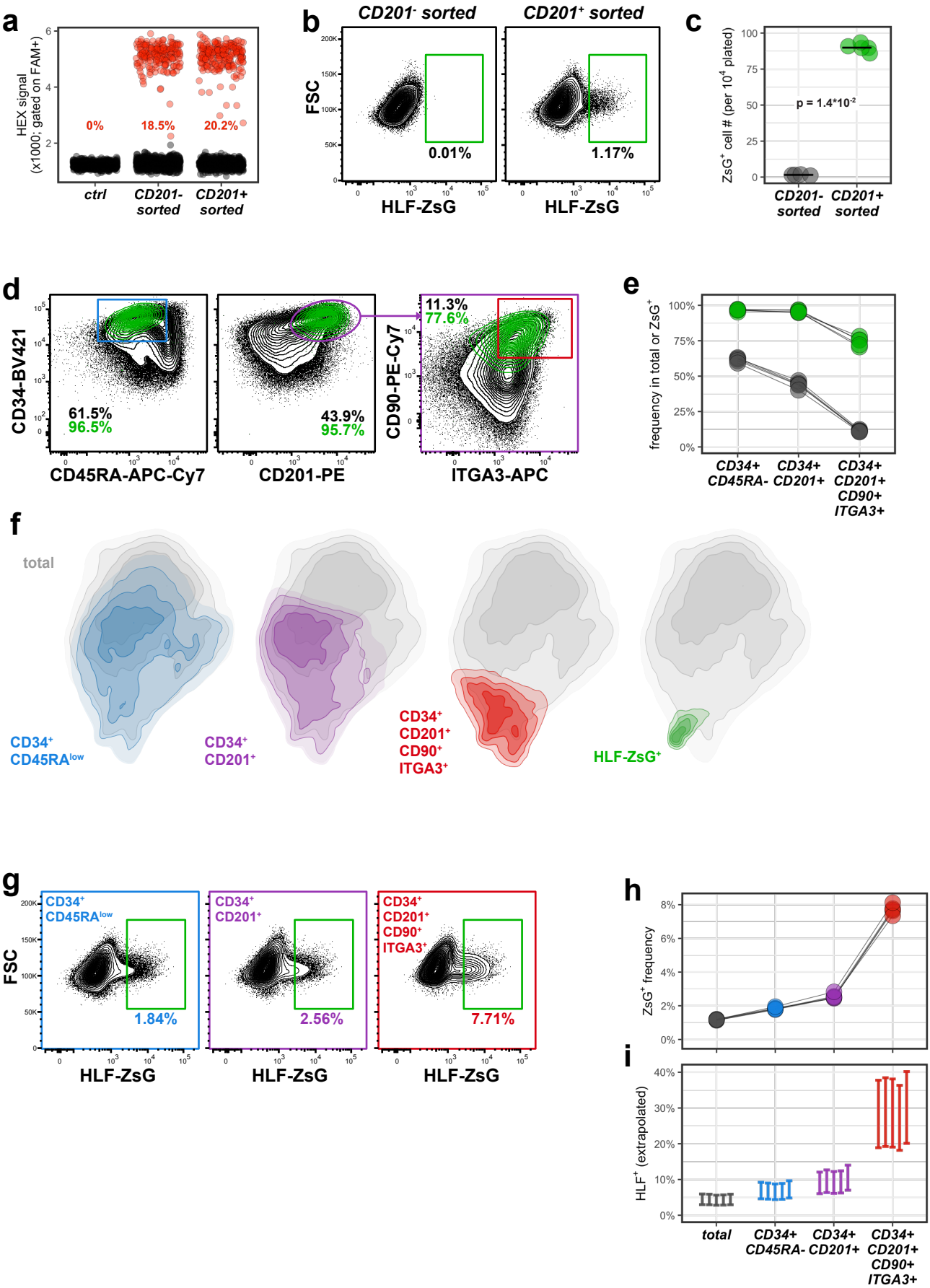

Fig. 4: *HLF* expression distinguishes repopulating cells in CD34+ cord blood cultures.

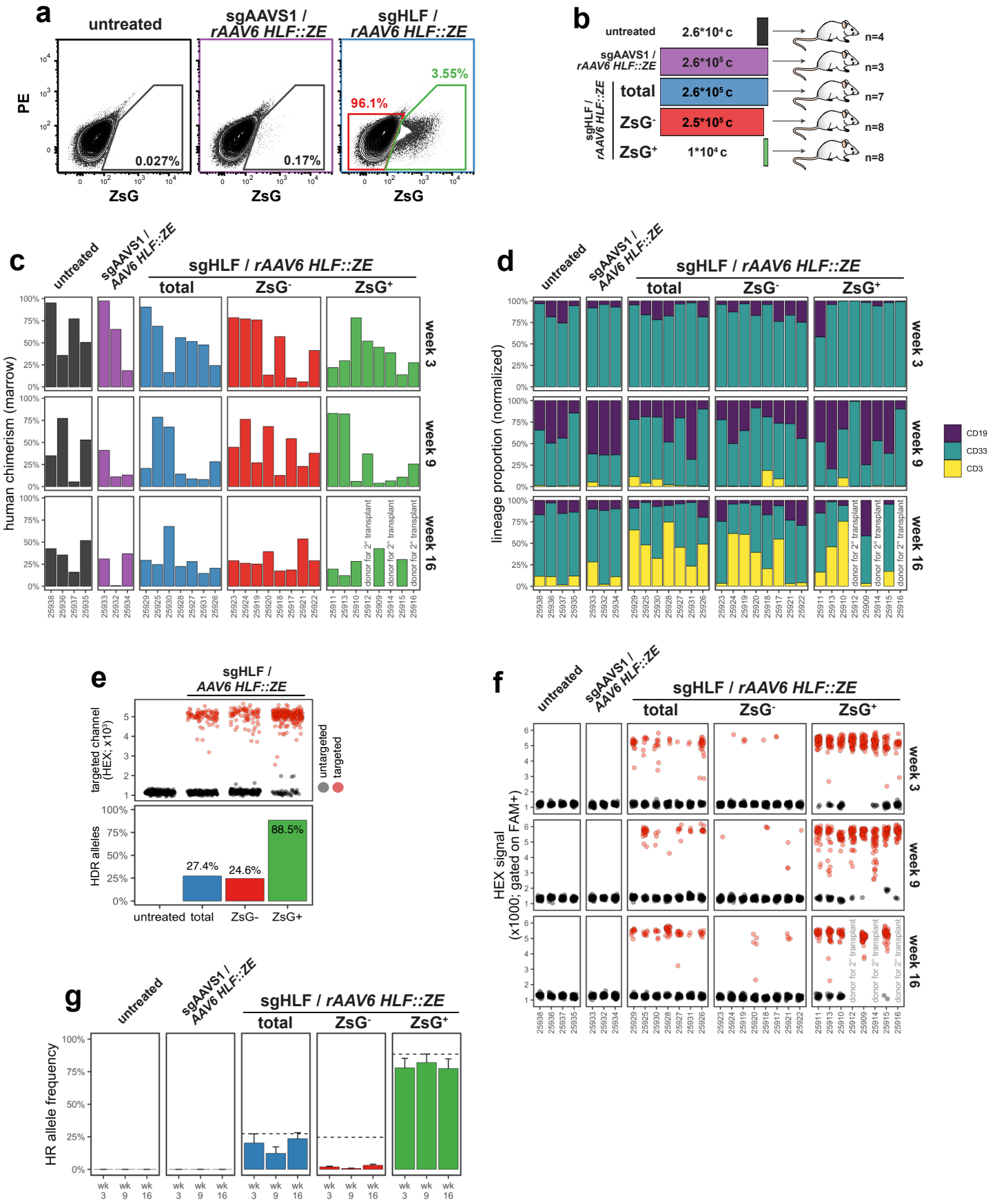
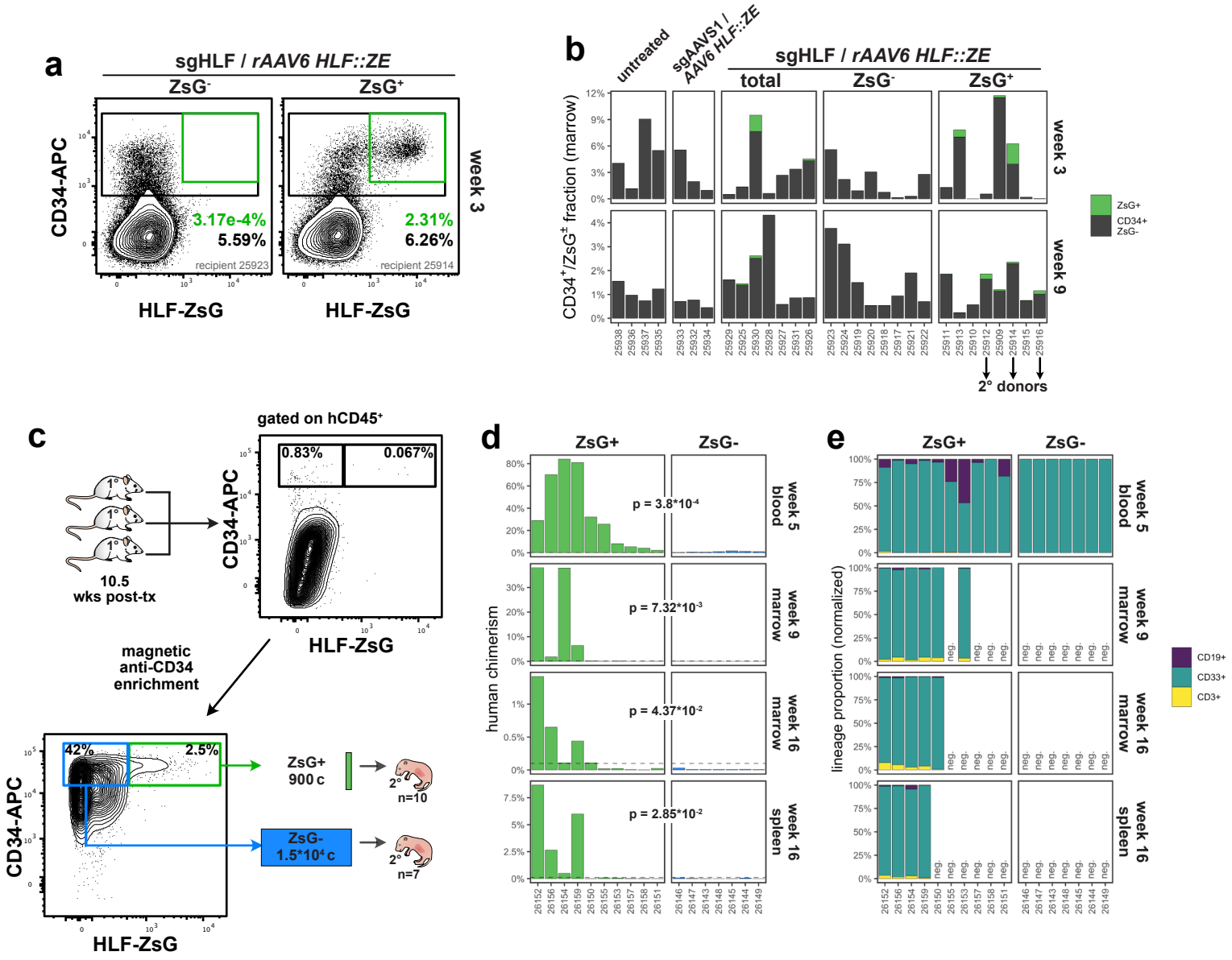
Fig. 5: *HLF*-expression labels HSCs with extensive self-renewal capacity.