nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-20792-y OPEN

Check for updates

The TRACE-Seq method tracks recombination alleles and identifies clonal reconstitution dynamics of gene targeted human hematopoietic stem cells

Rajiv Sharma[1,4], Daniel P. Dever[2,4], Ciaran M. Lee[3,4], Armon Azizi[1,4], Yidan Pan[3], Joab Camarena[2], Thomas Köhnke[1], Gang Bao[3✉], Matthew H. Porteus[2✉] & Ravindra Majeti[1✉]

Targeted DNA correction of disease-causing mutations in hematopoietic stem and progenitor cells (HSPCs) may enable the treatment of genetic diseases of the blood and immune system. It is now possible to correct mutations at high frequencies in HSPCs by combining CRISPR/Cas9 with homologous DNA donors. Because of the precision of gene correction, these approaches preclude clonal tracking of gene-targeted HSPCs. Here, we describe Tracking Recombination Alleles in Clonal Engraftment using sequencing (TRACE-Seq), a methodology that utilizes barcoded AAV6 donor template libraries, carrying in-frame silent mutations or semi-randomized nucleotides outside the coding region, to track the in vivo lineage contribution of gene-targeted HSPC clones. By targeting the *HBB* gene with an AAV6 donor template library consisting of ~20,000 possible unique exon 1 in-frame silent mutations, we track the hematopoietic reconstitution of *HBB* targeted myeloid-skewed, lymphoid-skewed, and balanced multi-lineage repopulating human HSPC clones in mice. We anticipate this methodology could potentially be used for HSPC clonal tracking of Cas9 RNP and AAV6-mediated gene targeting outcomes in translational and basic research settings.

[1] Department of Medicine, Division of Hematology, Cancer Institute, and Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, CA 94305, USA. [2] Department of Pediatrics, Stanford University, Stanford, CA 94305, USA. [3] Department of Bioengineering, Rice University, Houston, TX 77030, USA. [4] These authors contributed equally: Rajiv Sharma, Daniel P. Dever, Ciaran M. Lee, Armon Azizi. ✉email: gb20@rice.edu; mporteus@stanford.edu; rmajeti@stanford.edu

Genetic diseases of the blood and immune system, including the hemoglobinopathies and primary immunodeficiencies, affect millions of people worldwide with limited treatment options. Clinical development of ex vivo lentiviral (LV)-mediated gene addition in hematopoietic stem and progenitor cells (HSPCs) has demonstrated that a patient's own HSPCs can be modified and re-transplanted to restore proper cell function in the hematopoietic system[1]. While no severe adverse events have been reported resulting from insertional mutagenesis in more than 200 patients transplanted with LV ex vivo manipulated HSPCs[2], efficacy in restoring protein/cell function and ultimately disease amelioration has varied. In some diseases, this lack of therapeutic efficacy is possibly the result of irregular spatiotemporal transgene expression due to the semi-random integration patterns of LVs.

Tracking the transgene integration sites (IS) by deep sequencing has been used to "barcode" clones in heterogeneous cell populations that contribute to blood reconstitution in the human transplantation setting. In clinical trials, IS methodology has been used to track genetically modified memory T-cells[3], waves of hematopoietic repopulation kinetics[4], as well as dynamics and outputs of HSPC subpopulations in autologous graft composition[5]. These seminal studies provided new insights into the reconstitution of human hematopoiesis following autologous transplantation. Importantly, IS can also provide evidence of potential concerning integration patterns in tumor-suppressor genes, like *PTEN*[6], TET2[7], and *NF1*[8], which can be closely monitored during long-term follow-up to predict future severe adverse events.

Genetic barcoding on the DNA level has been used to track the in vitro[9] and in vivo[10–13] clonal dynamics of heterogeneous mammalian cellular populations and offers several advantages over lentiviral IS tracking, although it has not been used clinically. First, the amplified region is known and nearly the same for each barcode simplifying recovery from targeted cells, as opposed to semi-random LV integrations, which require amplification of unknown sequences. Second, it is far less likely for differences in amplification efficiency or secondary structure to lead to drop off or mis-quantification of clone sizes[14]. Altogether, genetic barcoding, combined with high-throughput sequencing, can enable sensitive and quantitative assessment of heterogeneous cell populations.

Genome editing provides an alternative approach to lentiviral integrations to perform permanent genetic engineering of cells. Genome editing can be performed using non-nuclease approaches[15,16], by base editing[17], or by prime-editing[18], but the most developed and efficient form of precision engineering in human cells utilizes engineered nuclease-based approaches[19–27]. The repurposing of the bacterial CRISPR/Cas9 system for use in human cells[28–31] has democratized the field of genome editing because of its ease of use, high activity, and high specificity, especially using high fidelity versions of Cas9[32]. Nuclease-based editing has now entered clinical trials with more on the horizon[33].

Genome editing by combining ribonucleoprotein (RNP, Cas9 protein complexed to synthetic stabilized, single guide RNAs) with the use of the non-integrating adeno-associated virus serotype 6 (AAV6) viral vector to deliver the donor template has been shown to be a highly effective system to modify therapeutically relevant primary human cells including HSPCs, T-cells, and induced pluripotent cells[34,35]. This approach has shown pre-clinical promise to usher in a new class of medicines for sickle cell disease[32,36], SCID-X1[37,38], MPS I[39], chronic granulomatous disease[40], X-linked Hyper IgM[41], and cancer[42]. The specificity of genome editing, however, means that with current approaches it is not possible to track the output of any specific gene-modified cell. The spectrum of non-homologous end joining (NHEJ)-introduced INDELs is also not broad enough to reliably measure clonal dynamics within a population[43]. Yet, understanding clonal dynamics within large populations of engineered cells is important and significant in both pre-clinical studies and potentially clinical studies. Therefore, we developed a barcode system for homologous recombination-based genome editing. We applied this system to understand the clonal dynamics of CD34$^+$ human HSPCs following transplantation into immunodeficient NSG mice.

In this study, we describe TRACE-Seq, a methodology that allows for both correction of disease-specific mutations and for the tracking of contributions of gene-targeted HSPCs to single and multi-lineage hematopoietic reconstitution. In brief, we demonstrate: (1) design and production of barcoded AAV6 donor templates using silent in-frame mutations or semi-randomized nucleotides outside the coding region (but inside the homology arms), (2) barcoding the first 9 amino acids of HBB exon 1 with ~20,000 possible AAV6 donor templates maintains high gene correction frequencies while preserving robust beta-globin expression levels, (3) the ability to track the reconstitution of gene-corrected myeloid- and lymphoid-skewed HSPC clones as well as balanced multi-lineage clones, and (4) an analysis pipeline that includes a highly adaptable platform for interpreting and summarizing rich datasets from clonal tracking studies that is deployable as a website accessible to researchers with no coding experience. TRACE-Seq demonstrates that Cas9 RNP and AAV6-mediated gene correction can be used to target a single HSC clone that can then robustly repopulate the myeloid and lymphoid branches of the hematopoietic system. This method and information further supports the translational potential of homologous recombination-based approaches for the treatment of genetic diseases of the blood and immune system.

Results

Design, production, and validation of barcoded AAV6 donor templates for targeting the HBB gene in human HSPCs. We previously developed an *HBB* AAV6 homologous donor template that corrects the sickle cell disease-causing mutation in HSPCs with high efficiencies[36]. Using this AAV6 donor as a template, we designed an *HBB* barcoded AAV6 donor library with the ability to: (1) correct the E6V sickle mutation, (2) preserve the reading frame of the beta-globin gene, and (3) generate enough sequence diversity to track cellular events on the clonal level (throughout the manuscript we will consider unique barcodes representative of cellular clones, with the caveat that clone counts may be overestimated due to bi-allelic targeting of two barcodes into the genome of a single cell). We designed the donor pool to contain mixed nucleotides that encode silent mutations within the first 9 amino acids of the HBB coding sequence ("VHLTPEEKS", Fig. 1a). Using this strategy, we designed double-stranded DNA oligos that contained the library of nucleotide sequences and cloned four separate pools of donors with a theoretical maximum number of 36,864 in-frame, synonymous mutations (Fig. 1b). This number of barcodes minimizes the potential of barcode collision (multiple long-term engrafting cells receiving the same barcode, and as a result erroneously considered to be related in lineage).

To ensure that the initial plasmid library reached the theoretical maximum diversity with near-equal representation of all sequences, we performed amplicon sequencing on the initial plasmid pools. Sequencing of HBB barcoded pools 1, 2, and 4 (Fig. 1a, bottom) revealed a wide distribution of sequences with no evidence of any highly overrepresented barcodes (Fig. 1c). Barcode pool 3 was eliminated for further study, because it was

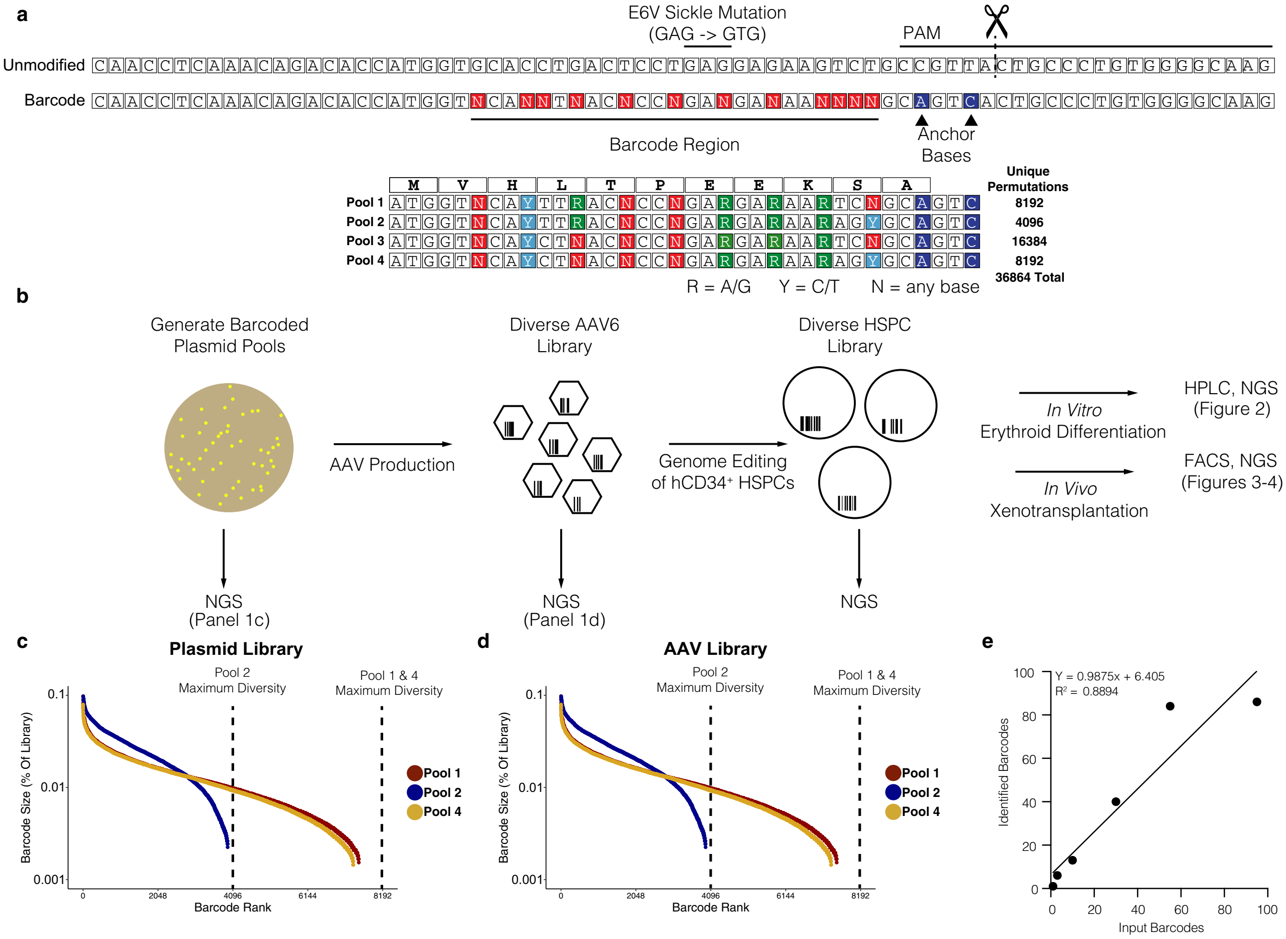
a
E6V Sickle Mutation
(GAG -> GTG)
PAM
Unmodified CAACCTCAAACAGACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAG
Barcode CAACCTCAAACAGACACCATGGTNCANNTNACNCCNGANGANAANNNNGCAGTCACTGCCCTGTGGGGCAAG
Barcode Region
Anchor Bases
M V H L T P E E K S A
Unique Permutations
Pool 1 ATGGTNCAYTTRACNCCNGARGARAARTCNGCAGTC 8192
Pool 2 ATGGTNCAYTTRACNCCNGARGARAARAGYGCAGTC 4096
Pool 3 ATGGTNCAYCTNACNCCNGARGARAARTCNGCAGTC 16384
Pool 4 ATGGTNCAYCTNACNCCNGARGARAARAGYGCAGTC 8192
36864 Total
R = A/G Y = C/T N = any base
b
Generate Barcoded Plasmid Pools
AAV Production
Diverse AAV6 Library
Genome Editing of hCD34⁺ HSPCs
Diverse HSPC Library
In Vitro Erythroid Differentiation
HPLC, NGS (Figure 2)
In Vivo Xenotransplantation
FACS, NGS (Figures 3-4)
NGS (Panel 1c)
NGS (Panel 1d)
NGS
c
Plasmid Library
Pool 2 Maximum Diversity
Pool 1 & 4 Maximum Diversity
Barcode Size (% Of Library)
0.1
0.01
0.001
0 2048 4096 6144 8192
Barcode Rank
Pool 1
Pool 2
Pool 4
d
AAV Library
Pool 2 Maximum Diversity
Pool 1 & 4 Maximum Diversity
Barcode Size (% Of Library)
0.1
0.01
0.001
0 2048 4096 6144 8192
Barcode Rank
Pool 1
Pool 2
Pool 4
e
Y = 0.9875x + 6.405
R² = 0.8894
Identified Barcodes
0 20 40 60 80 100
Input Barcodes
0 20 40 60 80 100

Fig. 1 Design and production of barcoded AAV6 donors for long-term genetic tracking of gene-targeted cells and their progeny. a Schematic of HBB targeting strategy. Top: Unmodified (WT) and barcoded HBB alleles depicted, with location of the E6V (GAG - > GTG) sickle cell disease mutation and CRISPR/Cas9 target sites labeled. Bottom: β-globin ORF translation with four barcode pools representing all possible silent mutations encoding amino acids 1-9. **b** Schematic of barcode library generation and experimental design. c/d Percentages of reads from each valid barcode identified through amplicon sequencing of plasmids (**c**) and AAV (**d**) pools 1, 2, and 4. **e** Recovery of barcodes from untreated genomic DNA containing 1, 3, 10, 30, and 95 individual plasmids containing HBB barcodes. Expected number of barcodes is plotted against the number of barcodes called by the TRACE-seq pipeline after filtering.

contaminated with uncut vector control and therefore skewed barcode diversity. After validating that the plasmid pools were diverse and lacked enrichment of any one sequence, we used the HBB barcoded library plasmid pools 1, 2, and 4 to produce libraries of AAV6 homologous donor templates. After generating barcoded AAV6 donor libraries, we performed amplicon-based NGS to determine the diversity and distribution of sequences. Similar patterns were observed, suggesting standard AAV6 production protocols do not introduce donor template bias in the barcoded pool (Fig. 1d).

Establishing thresholds for HBB barcode quantification. Understanding the clonal dynamics of hematopoietic reconstitution through sequencing requires the ability to differentiate between low frequency barcodes and noise introduced by sequencing error. Therefore, we used a modified version of the TUBAseq pipeline to cluster cellular barcodes and differentiate between sequencing error and bona-fide barcode sequences[44]. The overall schema of the TRACE-seq pipeline is depicted in Supplementary Fig. 1a. Briefly, we merged paired-end fastq files using the PEAR algorithm with standard parameters[45], and then aligned reads to the human HBB gene. Reads were binned into three categories: unmodified alleles (wildtype), non-homologous end joining (NHEJ) alleles (Supplementary Fig. 1a, b), and homologous recombination (HR) alleles. Reads were classified as unmodified if they aligned to the reference HBB gene with no genome edits. Reads were classified as NHEJ if there were any insertions or deletions within 20 bp of the cut site, and if anchor bases (PAM-associated bases changed after successful HR) were unmodified (Fig. 1a). Finally, reads were classified as HR if they had modified anchor bases and were not classified as NHEJ (Fig. 1a). All subsequent analyses were performed exclusively on the HR reads.

To differentiate between bona fide barcodes and sequencing errors, variable barcode regions, and non-variable training regions were extracted from the HR reads and TUBAseq was used to train an error model and cluster similar barcodes together using the DADA2 algorithm[44]. We chose a DADA2 clustering omega parameter of 10^{-40} because: 1) we found that at this omega value, the number of unfiltered barcodes called began to reach the minimum number of barcodes called per sample as omega was decreased, and 2) we found that varying this parameter did not ultimately affect the number or sequence of called barcodes after filtering (described subsequently) for samples with known barcode content (Supplementary Fig. 1c).

In order to benchmark our analysis pipeline, we cultured individual barcoded bacterial plasmid colonies in 96 well plates and generated pooled plasmid libraries to generate a set of ground-truth samples with known barcode content. These libraries were spiked into untreated human gDNA and were subjected to our optimized amplicon sequencing and analysis pipeline. We found that clustering eliminated more than 97% of low-level noise barcodes across all samples with known barcode content, but left a small percentage of low-level barcodes in the clustered barcode set (Supplementary Fig. 1d). Using the ground-truth samples, we determined a “high confidence” barcode threshold of 0.5%, which allowed us to quantitatively recover the expected numbers of barcodes ($R^2 = 0.89$) (Fig. 1e, Supplementary Fig. 1e). We performed subsampling analyses to establish the minimum numbers of input reads required to obtain reliable data through our pipeline (Supplementary Fig. 1f).

Overall, our pipeline allowed us to process raw amplicon sequencing data and generate a set of barcodes unlikely to contain spurious signals. Conceptually, we extracted barcodes from each read and eliminated barcodes which appeared to be derived from sequencing or other error using a clustering-based methodology and evidence-based filtering heuristics, resulting in a set of high-confidence barcodes with which we performed further analyses.

Barcoding HBB exon 1 with in-frame silent mutations preserves hemoglobin expression while allowing cell tracking within a heterogeneous population. To evaluate whether the barcoded AAV6 donor libraries preserved the open reading frame of HBB following targeted integration, we compared HSPCs targeted with a non-barcoded homologous donor (containing a single corrective AAV6 genome[36]; non-BC) or a barcode donor library (BC) as illustrated in Fig. 2a (see also Supplementary Table 1). We performed gene-targeting experiments by electroporating HiFi Cas9 and HBB-specific chemically modified guide sgRNAs[46] into primary CD34$^+$ HSPCs isolated from patients with sickle cell disease (which contained the E6V point mutation). We observed similar gene correction efficiencies between HSPCs targeted with non-BC and BC donors as quantified by amplicon-based next-generation sequencing from approximately 1000 cells from each timepoint (Fig. 2b). To assess barcode diversity, we ranked barcodes by read percentages from largest to smallest for each treatment group (Fig. 2c). Focusing specifically on the top 20 barcodes in the representative example in Fig. 2c, it is evident that even with a relatively small sample, we observe a fairly even distribution of barcodes, with no evidence of extreme overrepresentation from any particular sequences. We calculated the number of the most abundant barcodes comprising 50% and

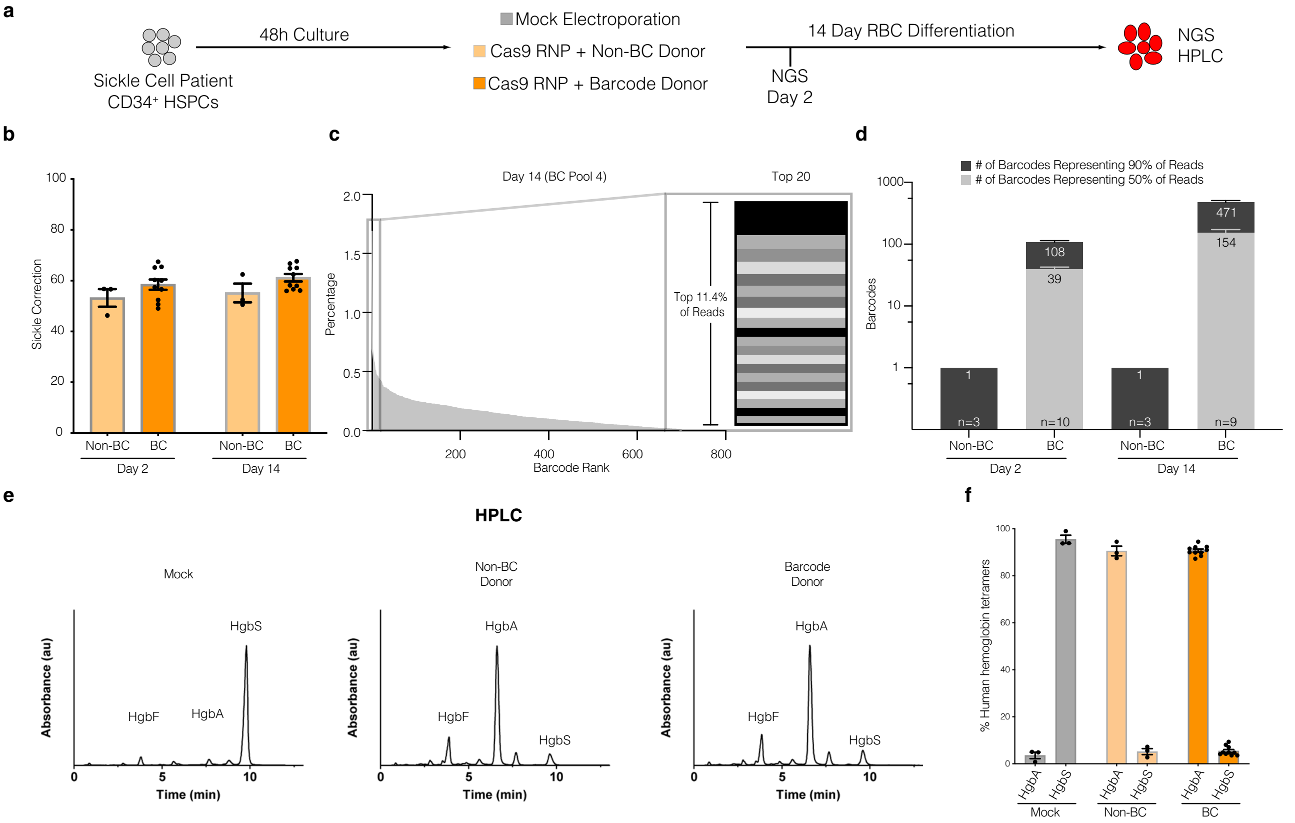

Fig. 2 Correction of the Sickle Cell Disease-causing E6V mutation using barcoded AAV6 donors in SCD-derived CD34$^+$ HSPCs. a Experimental design – SCD patient-derived CD34$^+$ HSPCs edited with CRISPR/Cas9 RNP and electroporation only (mock), single donor (non-BC), or barcode donor (BC) AAV6 HR templates. **b** SCD correction efficiency (percentage of corrected sickle cell alleles) of non-BC and BC treated groups as a fraction of total NGS reads (e.g., HR reads / [all reads].) $n = 3$ (non-BC) and $n = 10$ (BC) biologically independent replicates examined over 3 independent experiments. **c** Representative example of barcode fractions in descending order from one donor at day 14 time point. Right: Top 20 clones represented as stacked bar graph (representing 11.4% of reads). **d** Number of unique barcode alleles comprising the top 50% and top 90% of reads from each treatment condition, sampling approximately 1000 cells per condition (see Supplementary Tables 1 and 2). **e** Representative hemoglobin tetramer HPLC chromatograms of RBC differentiated cell lysates at day 14 post-treatment. **f** Quantification of total hemoglobin protein expression in each group. Each data point represents an individual biological replicate. $n = 3$ (Mock and non-BC) and $n = 10$ (BC) biologically independent replicates examined over 2 independent experiments. HgbA adult hemoglobin, HgF fetal hemoglobin, HbS sickle hemoglobin, AAV6 Recombinant AAV2/6 vector. Error bars depict mean ± SEM.

90% of total HR reads as a measure of sequence diversity. Note that these As expected, the single non-BC donor sample contained one barcode (the corrected E6V sequence) along with intentional synonymous mutations[36] that represented >94% of reads (Fig. 2d). Of note, the remaining reads appeared to be sequencing/PCR artifacts as they often contained nonsynonymous mutations in the HBB reading frame (Supplementary Table 2). In contrast, the 90th percentile of barcode reads in BC donor targeted cells contained a mean of 107.7 ± 9.6 barcodes at day 2 and 471.0 ± 54.1 at day 14 (Fig. 2d). Note that these counts reflect all barcodes, not only the "high confidence" <0.5%, in order to better represent the barcode distributions. These unique barcode counts were not surprising given the limited numbers of input cells analyzed, and the additional complexity of performing nested PCR reactions to avoid contamination from unintegrated (episomal) AAV6 donor genomes, especially at early timepoints before the cells could undergo many rounds of division. Indeed, by aggregating together all experimental replicates treated with BC donors, the 90th percentile of barcode reads contained >3200 barcodes suggesting barcode identification was limited by sampling depth. Importantly, the barcodes observed in the BC donor treated samples preserved the HBB coding sequence even though their sequences varied greatly (Supplementary Table 2). These results are consistent with the notion that targeting HSPCs with a BC donor produces a diverse pool of HSPCs capable of correcting the E6V sickle mutation, and that diversity is maintained within a two-week period of in vitro culture.

While the sequencing data suggest that the HSPCs targeted with the BC donors exhibit robust E6V gene correction frequencies, the introduction of silent mutations may interfere with hemoglobin protein expression. To assess this possibility, we performed in vitro erythroid differentiation of non-BC and BC targeted HSPCs and collected red blood cell pellets for HPLC analysis of hemoglobin tetramer formation. While the unedited mock sample contained >90% sickle hemoglobin (HgbS) (of total hemoglobin), HSPCs targeted with non-BC or BC AAV6 donors both exhibited >90% adult hemoglobin (HgbA) protein production (Fig. 2e-f). These results suggest the silent mutations introduced by the BC donor had no significant negative influence on overall translation efficiency, despite being produced from a diverse pool of >450 unique sequences in the bulk-edited population.

TRACE-Seq identifies long-term engraftment of lineage-specific and bi-lineage potent HBB targeted hematopoietic stem and progenitor cells. In addition to correcting the E6V mutation and restoring HgbA expression, barcoded AAV6 donors can be utilized to label and track cells in a heterogeneous pool of HSPCs. To track cellular lineages in a pool of HBB-labeled HSPCs, we transplanted BC and non-BC control targeted cord blood CD34$^+$ HSPCs via intra-femoral injection into sublethally irradiated adult female NSG recipient mice (2–4 × 10^5 cells per mouse from $n = 6$ total cord blood donors, see Supplementary Table 3). Upon sacrifice (16-18 weeks post-engraftment), mice in both transplantation groups exhibited no statistically significant differences in total human engraftment (46 ± 10.4 vs. 50 ± 10.1, non-BC and BC, respectively, Fig. 3a). Similarly, no significant differences were seen between non-BC and BC mice in terms of lineage reconstitution of the human cells engrafted, which mainly consisted of B cells (CD19$^+$), myeloid cells (CD33$^+$), or HSPCs (CD19$^-$CD33$^-$CD10$^-$CD34$^+$) (Fig. 3b).

To evaluate the efficiency of non-BC or BC gene targeting in long-term engrafting HSPCs, bone marrow MNCs were sorted by flow cytometry into lineages CD19$^+$ and CD33$^+$, as well as the multipotent HSPC (CD19$^-$CD33$^-$CD10$^-$CD34$^+$) populations. Using the pipeline outlined in Supplementary Fig. 1a, we performed amplicon-based NGS to quantify the proportions of gene-targeted alleles relative to total editing events that included NHEJ and unmodified alleles. To further validate our threshold for "high confidence" barcodes, we performed in silico mixing simulations as described in Supplementary Fig. 1e, which resulted in very few false positive or false negative barcode calls at our previously determined threshold of 0.5%. Importantly, the value of the threshold did not have a major impact on the median numbers of barcodes called (Supplementary Fig. 1g). We did not detect any significant differences in the efficiency of HR within any of these subsets between non-BC and BC donors (Fig. 3c).

Because there was robust engraftment of HBB targeted alleles in the BC mice, we were able to track the recombination alleles within the lymphoid, myeloid, and multipotent HSPC subpopulations. We analyzed cells from a total of 9 mice sorted on lymphoid (CD19$^+$), myeloid (CD33$^+$), and HSPC (CD19$^-$CD33$^-$CD10$^-$ CD34$^+$) markers. 130.6 ± 62.3 unique barcodes accounted for 90% of the reads with a median of 2 unique barcodes accounting for 50% or the sequencing reads from each group (Fig. 3d). As in Fig. 2d, the counts by percentile reflect all barcodes called through the TRACE-seq pipeline. Barcodes in all three sorted populations (Supplementary Fig. 2 and Supplementary Table 4) exhibited less diversity than was observed in vitro, indicating that there was a reduction in clonal complexity following engraftment into mice (Supplementary Fig. 3a). For example, the CD19$^+$ compartment from Mouse 18 contained over 60 total clones passing our thresholds, with a majority of reads coming from a single barcode (Supplementary Fig. 3b). The number of high confidence barcodes (>0.5% of reads) was correlated with total human engraftment in the lymphoid compartment and a similar trend was observed in the myeloid compartment ($p = 0.08$) (Fig. 3e). The same trend was observed when we correlated barcodes with lineage-specific engraftment adjusted for HR frequency (Fig. 3f). When we subdivided these more abundant barcodes into alleles that contributed to lymphoid only, myeloid only, or bi-lineage output within the mice, we observed fewer barcodes generated from lymphoid-skewed compared to myeloid-skewed or bi-lineage HSPCs ($p = 0.0013$ and $p = 0.024$, respectively, Fig. 3g). These data suggest that Cas9/sgRNA and AAV6-mediated HBB gene targeting occurs in multipotent HSPCs and/or lineage-restricted HSPCs.

The gold standard for defining human long-term hematopoietic stem cell (LT-HSC) activity is to perform secondary transplants into another sublethally irradiated NSG mouse[47]. Therefore, we compared the TRACE-Seq dynamics of a primary recipient versus a secondary recipient in mouse 20 that exhibited very high engraftment (>80% human cell engraftment). While mouse 20 had a total of 17 lymphoid and 56 myeloid clones contributing to the engraftment of gene-targeted HBB cells, the majority of differentiated cell output was from relatively few clones (Fig. 4a, left panel). Four lymphoid and five myeloid lineage barcodes accounted for 50% of the reads from each population. This trend was consistent between all mice analyzed (Supplementary Fig. 3c–k) with each mouse displaying a unique set of HBB barcodes that all maintained the coding region (Supplementary Table 5). Barcode reads from the same sorted cell populations from the secondary mouse transplant revealed further reductions in clonal diversity, almost to a monoclonal state, with a single clone representing 80% or more of reads in both lymphoid and myeloid lineages (Fig. 4a, right panel, dark blue). Interestingly, the dominant clone in the secondary transplant was not the most abundant clone in the primary mouse as it only represented 10.9% of lymphoid and 16% of myeloid alleles.

To understand the contribution of each clone to the absolute number of differentiated hematopoietic cells in the mouse bone

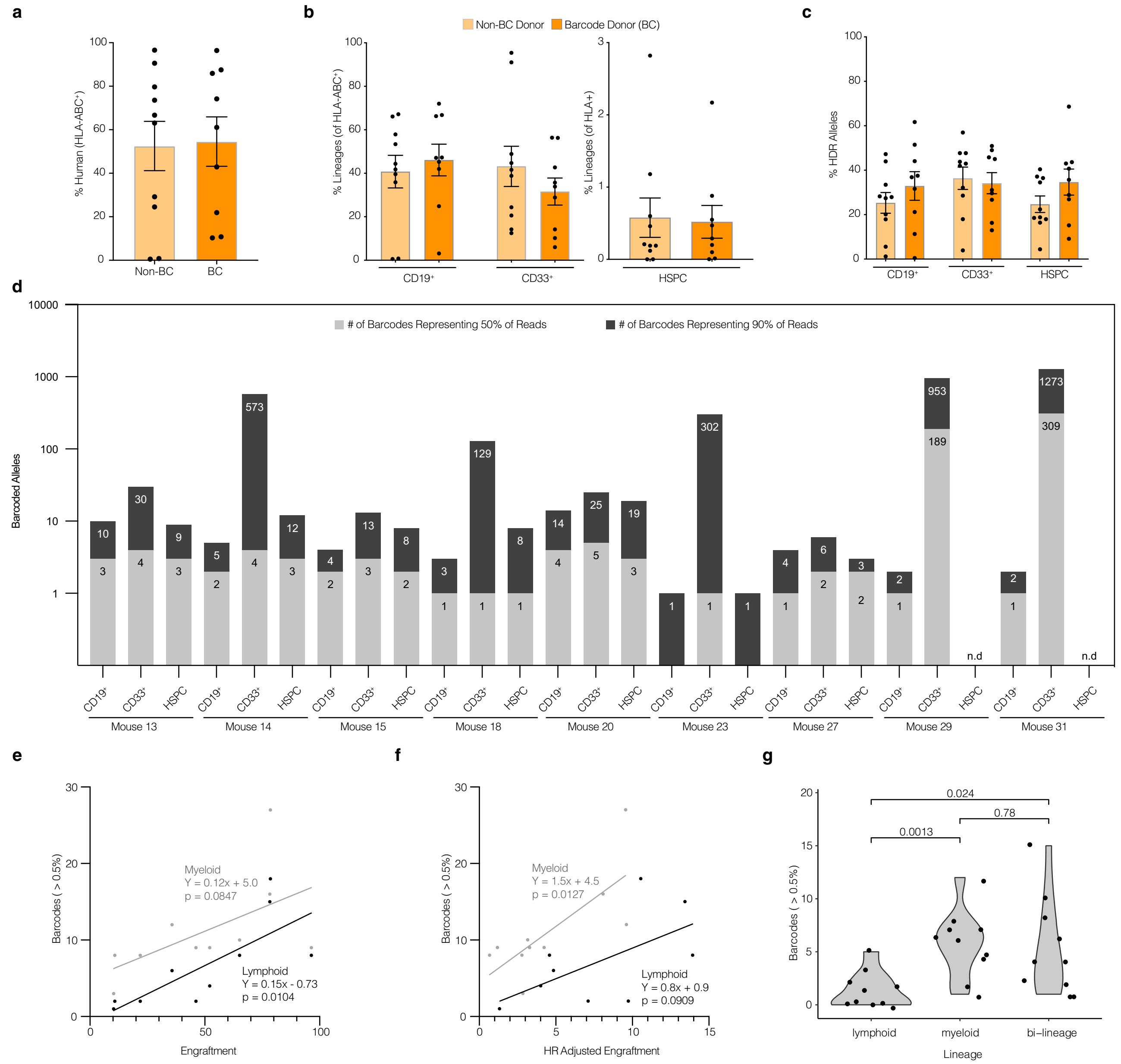

Fig. 3 TRACE-Seq identifies lineage-restricted and multi-potent gene-targeted HSPCs in primary NSG transplants. CD34$^+$ enriched cord blood-derived HSPCs were cultured in HSPC media containing SCF, FLT3L, TPO, IL-6, and UM-171 for 48 h, electroporated with Cas9 RNP (HBB sgRNA), transduced with AAV6 donors (either BC or non-BC), and cultured for an additional 48 h prior to intrafemoral transplant into sublethally irradiated NSG mice (total manufacturing time was less than 96 h). A total of 16–18 weeks post-transplantation, total BM was collected and analyzed for engraftment by flow cytometry, sorted on lineage markers, and sequenced for unique barcodes. Two independent experiments were performed to assess reproducibility of identifying clonality of gene-targeted HSPCs. **a** Total human engraftment in whole bone marrow, (as measured by the proportion of human HLA-ABC$^+$ cells). **b** Multilineage engraftment of human CD19$^+$, CD33$^+$, and HSPCs (CD19$^-$CD33$^-$CD10$^-$CD34$^+$). **c** Genome editing efficiency in each indicated sorted human lineage subset as determined by NGS (HR reads/[all reads]). **d** Barcodes from each subset were sorted from largest to smallest by percentage of reads. Depicted are the numbers of most abundant, unique barcode alleles comprising the top 50% and top 90% of reads from each lineage of all mice transplanted with BC donor edited HSPCs. Mean ± SEM genomes analyzed from each group: CD19$^+$: 8500 ± 1000, CD33$^+$: 8800 ± 800, HSPC: 1500 ± 500 (see Supplementary Table 4). **e** Correlation between numbers of high confidence barcodes (>0.5%) in lymphoid (gray) and myeloid (black) compartments and total human engraftment (as percent of human and mouse BM-MNCs). Lymphoid and myeloid values plotted for $n = 9$ primary engrafted mice and $n = 1$ secondary engrafted mouse. **f** Correlation between numbers of high confidence barcodes (>0.5%) in lymphoid (gray) and myeloid (black) compartments and HR adjusted engraftment ([human engraftment] x [lineage specific engraftment] x [HR efficiency]). Lymphoid and myeloid values plotted for $n = 9$ primary engrafted mice and $n = 1$ secondary engrafted mouse. **g** Numbers of high confidence barcodes from each mouse which contribute to lymphoid only (CD19$^+$), myeloid only (CD33$^+$), or both lineages. High confidence barcodes: barcodes with at least 0.5% representation (see Supplementary Fig. 1). All points represent individual mice ($n = 9$ non-BC treated and $n = 10$ BC treated biologically independent mice in one independent experiment), with the exception of panels **e**–**g** (where barcodes from each mouse are separated based on lineage contribution). Error bars depict mean ± SEM. p values reflect 2-tailed t-test.

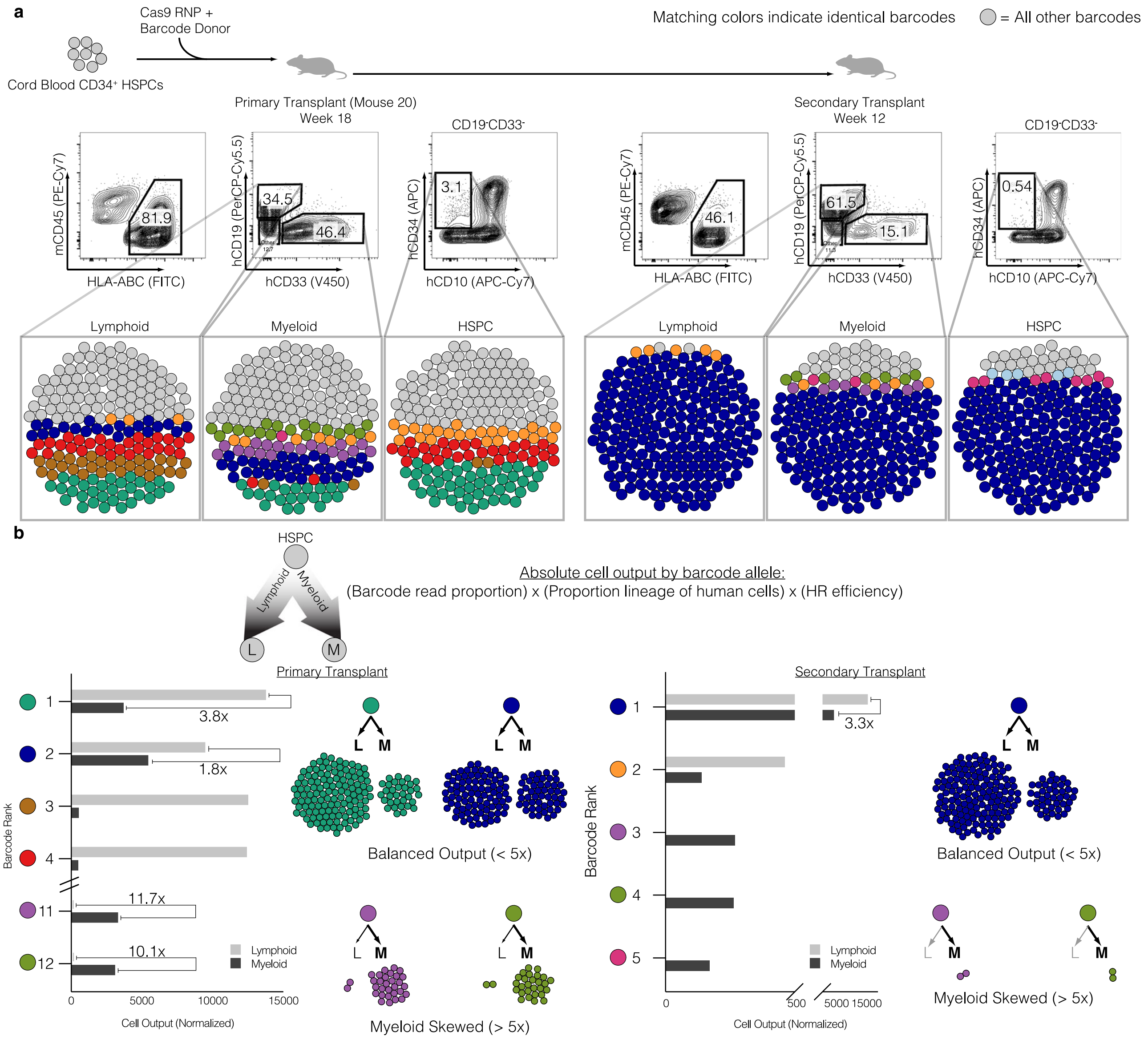

Fig. 4 Identification of clonal dynamics of HBB-targeted HSPCs. a Top: Experimental schematic. Middle: Flow cytometry plots representing robust bi-lineage engraftment in primary transplant (left, week 18 post-transplant) and secondary transplant (right, week 12 post-transplant). Bottom: Bubble plots representing barcode alleles as unique colors from each indicated sorted population. Shown are the three most abundant clones from all six populations. All other barcodes represented as gray bubbles. **b** Normalized output of barcode alleles with respect to lineage contribution. Total cell output (bar graphs) from indicated barcodes adjusted for both differential lineage output and genome editing efficiency within each subset. Examples of various lineage skewing depicted, with cell counts proportional to the absolute contribution to the xenograft. Skewed output defined as 5-fold or greater bias in absolute cell counts towards lymphoid or myeloid lineages.

marrow, we took the product of the following parameters: 1) the fraction of unique barcode reads assigned to each clone, 2) the proportion of the lineage (CD19 or CD33) to total human engraftment within the mouse, and 3) the gene targeting efficiency (HR%) of each lineage (Fig. 4b). This analysis reveals clones that are lymphoid skewed (brown and red, Fig. 4a), myeloid skewed (purple and light green), as well as clones exhibiting balanced hematopoiesis (dark blue). We defined skewing as having a >5-fold difference in proportion between lymphoid and myeloid cells. Perhaps the most interesting observation from this analysis was that the more balanced hematopoietic clone (dark blue) was responsible for a great majority of secondary engraftment/repopulation (Fig. 4b, right). Interestingly, while this clone contributed >80% of the engraftment of HBB targeted cells, there were still observable myeloid lineage-skewed clones present in the secondary transplant. This analysis also revealed barcode sequences that produced highly correlated read frequencies (±2% read proportions) in both primary and secondary transplants, consistent with bi-allelic gene targeting in the same long-term HSPC (Fig. 4b, purple and light green barcodes).

TRACE-Seq by barcoding AAV6 donor templates outside the coding region allows for clonal tracking of AAVS1 targeted HSPCs. To test that the barcoding scheme (inside the coding region), library diversity (maximal theoretical diversity of 36,864 HBB barcodes), and/or the gene being targeted (HBB) did not bias our results, we developed a strategy to target *AAVS1* with barcoded SFFV-BFP-PolyA AAV6 donor libraries (Supplementary

Fig. 4a). We designed the AAVS1 barcoded variable region within the 3' untranslated region of the BFP expression cassette so the barcode would be in the genomic DNA as well as mRNA. Using a design that prevents mononucleotide runs that can potentially increase sequencing error[48], a 12 nucleotide variable barcode region resulted in a theoretical maximal barcoded AAV6 pool of 531,441 different homologous donor templates (Supplementary Fig. 4a, bottom). Using such a large pool allowed us to rule out the possibility that the numbers of barcodes observed in the HBB system is artificially limited by the smaller diversity of the HBB barcode pool. As with the HBB pipeline (Fig. 1e), we benchmarked our ability to differentiate sequencing error from legitimate barcodes by choosing parameters (e.g., barcode threshold 0.5%) that resulted in a high correlation between known numbers of input barcodes and barcodes identified through TRACE-seq (Supplementary Fig. 4b).

We targeted cord blood-derived HSPCs with the AAVS1-BC pool of AAV6 donor templates and transplanted them into sublethally irradiated NSG mice to assess the clonal contribution via TRACE-Seq. Robust AAVS1-BC donor targeting into the *AAVS1* locus was achieved in two independent experiments across five HSPC donors and a mean of $2.90 \pm 0.4 \times 10^5$ cells transplanted per NSG mouse (Supplementary Table 3). Following 16–18 weeks of hematopoietic reconstitution, we observed 45.4% ± 14.2 human engraftment, with a gene targeting efficiency of 42.4% ± 11.4 (Supplementary Fig. 4c). As with the HBB donors, the majority of differentiated cells were $CD19^+$ lymphoid and $CD33^+$ myeloid cells, with a strong trend towards more genome editing within the $CD33^+$ population (55.8 ± 12.0 vs. 22.3 ± 11.2; $p = 0.06$, two-tailed t-test) (Supplementary Fig. 4d, e). To assess clonal contributions of AAVS1 targeted HSPCs, lineage-specific cells ($CD19^+$ or $CD33^+$) were sorted (Supplementary Fig. 4e), and AAVS1-BFP specific amplicons were generated for NGS sequencing of cells with on-target integrations of SFFV-BFP-PolyA. Consistent with our findings targeting the *HBB* locus, we identified not only similar numbers of unique barcodes (representing individual clones) in divergent hematopoietic lineages (Fig. 5 a; Supplementary Fig. 4f), but also similar patterns between primary and secondary transplants, suggesting again that TRACE-Seq identifies Cas9/sgRNA and AAV6-mediated targeting of LT-HSCs (Fig. 5a, Supplementary Fig. 5). Across all mice, bi-lineage clones were seen in four out of five mice, with the exception being mouse 38, from which we were not able to sort sufficient numbers of myeloid cells for valid analysis (Supplementary Fig. 5 and Supplementary Tables 6 and 7). As with HBB TRACE-Seq, calculating the relative cell output of individual barcodes revealed lymphoid skewed, myeloid skewed, and balanced HSPC clones (Fig. 5b, left). The most dominant clone (red), which displayed high proliferative output with a more balanced hematopoietic lineage distribution in the primary mouse, was the predominant clone in the secondary transplant (Fig. 5b, right). Interestingly, this clone was not very abundant in the HSPC fraction of the primary transplant (<0.5%). In addition, we observed less abundant, myeloid skewed clones (blue and green) in both primary and secondary transplants. These results confirm that gene-targeted LT-HSC clones contribute to robust multi-lineage engraftment. Importantly, in both targeted loci, the large numbers of input barcodes relative to indels enabled us to have much more confidence that matching sequencing reads were likely to be a product of the same progenitor cell (Supplementary Data 1).

Discussion

TRACE-Seq improves the understanding of the clonal dynamics of hematopoietic stem and progenitor cells following homologous recombination-based genome editing using two different gene targets (*HBB* and *AAVS1*). The data demonstrate that Cas9/sgRNA and AAV6 gene editing targets four distinct types of hematopoietic cells capable of engraftment, including: (1) rare and potent hematopoietic balanced LT-HSCs, (2) rare lymphoid skewed progenitors, (3) rare and potent myeloid skewed progenitors, and (4) more common and less proliferative myeloid skewed HSPCs.

TRACE-seq clearly demonstrates that in the NSG mouse model, engraftment of human cells after genome editing is largely oligoclonal with a few clones contributing to the bulk of hematopoiesis. From a technical perspective, we have developed a data analysis pipeline with multiple filters to distinguish sequencing artifacts from low abundance clones. Although our study was limited in its ability to quantify very low abundance barcodes, improvements in sequencing platforms and barcode design (such as error-correcting barcodes), the ability to distinguish noise from low abundance clones will enable more sensitive barcode detection in the near future. Nonetheless, the evidence that clones that were seemingly rare in primary transplants can contribute significantly to hematopoiesis in secondary transplants demonstrates both the sensitivity of this method to detect such clones and the biologic importance of such clones in hematopoiesis.

We compare and contrast these results to lentiviral-based genetic engineering of HSPCs since clonal dynamics of genome-edited cells have not been published previously. Previous studies tracking LV IS in NSG mice have suggested on the order of 10 −200 total clones (without data regarding the relative contributions of different clones) persisting long-term (although at different frequencies in each of the two mice analyzed), with identification of lineage-skewed as well as multi-potent LT-HSCs[49]. Accordingly, TRACE-Seq identified >50 clones per mouse that were contributing to the entire hematopoiesis of gene-targeted cells (Fig. 3e), suggesting that genome-edited human HSPCs engraft as efficiently as lentiviral engineered cells in the NSG xenogeneic model. Interestingly, we identified 1–3 clones capable of robust multi-lineage reconstitution in secondary transplants, suggesting between one in 6×10^4 and 4.6×10^5 input cells are gene-targeted LT-HSCs (based on the numbers of cells transplanted). In a clinical trial for Wiskott-Aldrich syndrome (WAS), IS analysis showed the frequency of $CD34^+$ HSPCs with steady-state long term lineage reconstitutions falls between 1 in 100,000 and 1 in a 1,000,000 (a few thousand clones out of the ~80–200 million HSPCs transplanted)[4]. Further building on this clinical trial, recent reports have suggested that LV integrations occur in cells within the HSPC pool that have long-term lymphoid or myeloid lineage restrictions as well[5]. Taken together, our data suggest that the frequency of gene-targeting and LV gene addition are similar in potent long-term engrafting LT-HSCs.

TRACE-Seq also demonstrated genome-edited clones that were heavily lineage skewed in both primary and secondary transplants. This finding demonstrates that the gold standard of HSC function, namely serial transplantation, may not always identify multi-potent HSCs. Nonetheless, the method should allow assessment of other mouse xenograft models of human hematopoietic transplantation in supporting lineage-restricted and multi-lineage reconstitution of genome-edited cells, including models that further maintain healthy and leukemic myeloid and innate immune system development[50–52]. In the future, this method, potentially combined with cell sorting schemes to resolve lineage preference within the CD34 + fraction[53], should help determine whether cells that undergo gene targeting have a bias towards particular lineages which may help guide which human genetic diseases of the blood may be most amenable to gene targeting based approaches. For example, if gene targeting preferentially occurs in long-term myeloid progenitors, this would support its use in diseases that require long-term myeloid engraftment of

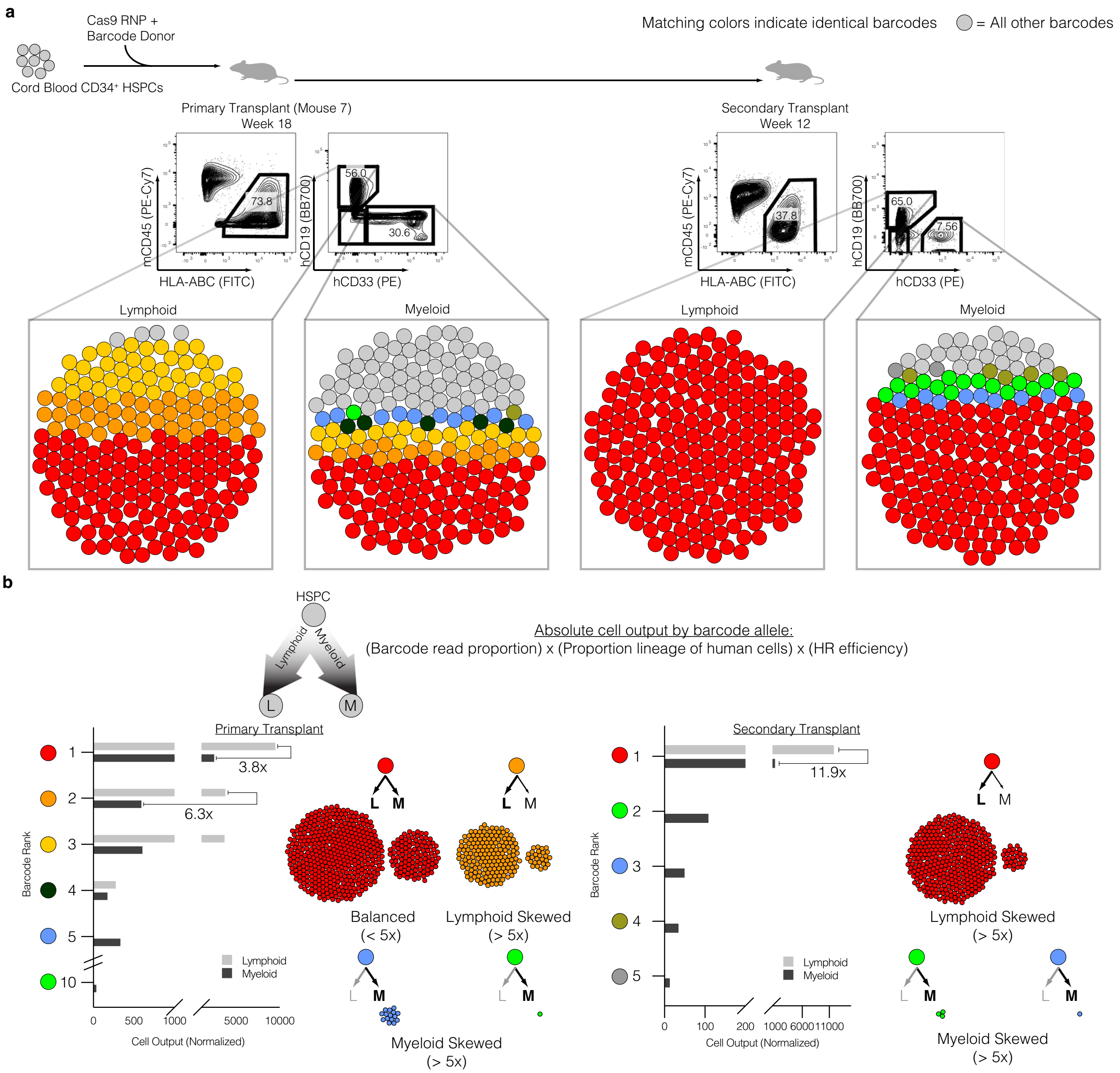

Fig. 5 Clonal tracking of AAVS1 barcoded targeted HSPCs in reconstituting primary and secondary NSG transplants. a Top: Experimental schematic. Middle: Flow cytometry plots representing bi-lineage engraftment in primary transplant (left, week 18 post-transplant) and secondary transplant (right, week 12 post-transplant). Bottom: Bubble plots representing barcode alleles as unique colors from each indicated sorted population. Shown are the three most abundant clones from all six populations. All other barcodes represented as gray bubbles. **b** Normalized output of barcode alleles with respect to lineage contribution. Total cell output (bar graphs) from indicated barcodes adjusted for both differential lineage output and genome editing efficiency within each subset. Examples of various lineage skewing depicted, with cell counts reflecting relative contributions to the xenograft. One highly engrafted mouse (Mouse 7) depicted of *n* = 5 total. Skewed output defined as 5-fold or greater bias in absolute cell counts towards lymphoid or myeloid lineages.

gene-targeted cells such as sickle cell disease, chronic granulomatous disease, or beta-thalassemia.

In addition to helping understand hematopoietic reconstitution of genome-edited cells in pre-clinical models, TRACE-Seq could also be used to further investigate the wide variety of genome editing approaches and HSC culture conditions to determine if they change either the degree of polyclonality or the lineage restriction of clones following engraftment. The wide numbers of variables that are under active study include different genome editing reagents and methods (different nucleases and donor templates and the inhibition of certain pathways[54,55]), differing culture conditions (e.g., cytokine variations[56], small molecules[57,58], peptides[54,55], and 3-D hydrogel scaffolds[59]), and altering the metabolic or cell cycle properties of the gene-edited cells. This study, in which two different approaches targeting two different genes were established, serves as the foundation for such future studies.

In conclusion, TRACE-seq demonstrates that homologous recombination-based genome editing can occur in human hematopoietic stem cells as defined by multi-lineage reconstitution following serial transplantation at a single cell, clonal level. Moreover, TRACE-Seq lays the foundation of clonal tracking of gene-targeted HSPCs for basic research into normal and malignant hematopoiesis. The ability of track clones in a clinical setting has proven to be a powerful approach to understand the safety, efficacy, and clonal dynamics of lentiviral-based gene therapies, and it will be informative to determine if regulatory agencies will accept having innocuous barcodes as part of recombination donor templates in clinical studies so that the safety, efficacy, and clonal dynamics of reconstituted gene-targeted cells, including HSCs, T-cells, or other engineered cell types, can be tracked following administration to patients.

Methods

Donor design and cloning

HBB barcode donor libraries. AAV transfer plasmid with inverted terminal repeats (ITR) from AAV2 that contained 2.4 kb of the *HBB* gene previously described[36] was digested with NcoI and BamH1 restriction enzymes (NEB) that resulted in deletion of a 435 bp band and the digested backbone was collected for further subcloning. Double-stranded DNA gBlock (IDT) pools with degenerate bases representing silent mutations containing 645 bases of homology were ordered in four separate oligo pools (as detailed below with bold depicting silent mutation region). Four different barcoded dsDNA oligo pools were ordered to maximize potential silent mutations that if all were ordered in the same library would have resulted in amino acid changes to the coding region. Each HBB barcoded dsDNA pool was then digested with NcoI and BamHI resulting in a 435 bp band that was collected and purified. NEB Assembly ligation reactions were performed for 1 h at 50 °C using a digested, gel-purified vector. Ligated HBB barcoded donor pools were transformed using NEB DH10B electrocompetent bacteria (NEB C3020K) or XL10-Gold competent cells (Agilent 200315) according to the manufacturer's protocol. At least two times the theoretical maximum number of possible barcoded donor templates were plated to ensure the generation of as much diversity as possible. Endotoxin-free maxipreps were generated for AAV6 production and purification. As noted, *HBB* barcode pool 3 was not included in genome editing experiments because enrichment of the original undigested donor plasmid was seen during sequencing of the plasmid library.

HBB barcode pool 1 (8192 possible unique donor templates):
atgg**TNCAYTTRACNCCNGARGARAARTCNGCAGTCACT**gcc
HBB barcode pool 2 (4096 possible unique donor templates):
atgg**TNCAYTTRACNCCNGARGARAARAGYGCAGTCACT**gcc
HBB barcode pool 3 (16384 possible unique donor templates):
atgg**TNCAYCTNACNCCNGARGARAARTCNGCAGTCACT**gcc
HBB barcode 4 (8192 possible unique donor templates):
atgg**TNCAYCTNACNCCNGARGARAARAGYGCAGTCACT**gcc

AAVS1 barcode donor libraries. AAVS1 barcode libraries were generated similarly to HBB libraries. Briefly, degenerate nucleotides (following the pattern "VHDBVHDBVHDB," in order to minimize homopolymer stretches[48]) were introduced by PCR 3' of the mTagBFP2 reporter cassette. pAAV-MCS plasmid (Agilent Technologies) containing ITRs from AAV serotype 2 (AAV2) was digested with NotI and barcode-containing PCR fragments were assembled into the backbone using NEB Assembly using the following primers, prior to transformation with XL10-Gold competent cells (Agilent 200315):

Insert_Fw1: CCATCACTAGGGGTTCCTGCGGCCGCCACCGTTTTTCT
Insert_Rv1: TTAATTAAGCTTGTGCCCCAGTTTGCTAGG
Insert_Fw2: TGGGGCACAAGCTTAATTAA**VHDBVHDBVHDB**CTCGAGG
GCGC
Insert_Rv2: CCATCACTAGGGGTTCCTGCGGCCGCAGAACTCAGGAC.

AAV6 production and purification. HBB barcoded recombinant adeno-associated virus serotype six (AAV6) vectors were produced and purified as previously described[60]. Briefly, 293FT cells (Life Technologies) were seeded at 15 million cells per dish in a total of ten 15-cm dishes one to two days before transfection (or until they are 80–90% confluent). One 15-cm dish was transfected with 6 µg ITR-containing *HBB* barcoded donor plasmid pools 1–4 and 22 µg pDGM6. Cells were incubated for 48–72 h until collection of AAV6 from cells by three freezes-thaw cycles. AAV6 vectors were purified on an iodixanol density gradient, AAV6 vectors were extracted at the 60–40% iodixanol interface, ßßand dialyzed in PBS with 5% sorbitol with 10 K MWCO Slide-A-Lyzer G2 Dialysis Cassette (Thermo Fisher Scientific). Finally, vectors were added to pluronic acid to a final concentration of 0.001%, aliquoted, and stored at -80 °C until use. AAV6 vectors were tittered using digital droplet PCR to measure the number of vector genomes using primers landing in the ITR regions[61]. AAVS1 barcoded AAV6 donors were produced as described above but purified using a commercial purification kit (Takara Bio #6666).

$CD34^+$ hematopoietic stem and progenitor cell culture. All $CD34^+$ cells used in these experiments were cultured as previously described[60]. In brief, cells were cultured in low-density conditions (<250,000 cells/mL), low oxygen conditions (5% O_2), in SFEMII (Stemcell Technologies) or SCGM (CellGenix) base media supplemented with 100 ng/mL of TPO, SCF, FLT3L, IL-6 and the small molecule UM-171 (35 nM). For in vitro studies presented in Fig. 2, $CD34^+$ cells from sickle cell disease patients were obtained as a kind gift from Dr. John Tisdale at the National Institute of Health (that were mobilized with plerixafor in accordance with their informed consent) or from routine non-mobilized peripheral blood transfusions at Stanford University under informed consent. For in vivo studies presented, cord blood-derived $CD34^+$ cells were purchased from AllCells or STEMCELL Technologies and were thawed according to the manufacturer's recommendations.

Cas9/sgRNA and AAV6-mediated genome editing. All experiments in these studies used the R691A HiFi Cas9 mutant[32] (IDT and Aldevron), and chemically synthesized guide RNA (sgRNA)[46] (Synthego). The guide sequences were as follows: HBB: 5′-CTTGCCCCACAGGGCAGTAA-3′ and AAVS1: 5′-GGGGCCAC TAGGGACAGGAT-3′. Genome editing experiments using Cas9/sgRNA and AAV6 were performed similarly to previously described methods[60]. In brief, $CD34^+$ HSPCs were thawed and plated for 48 h to allow for recovery of freezing process and pre-stimulation of cell cycle. $CD34^+$ HSPCs were then electroporated in 100 µl electroporation reaction buffer P3 (Lonza) with 30 µg HiFi (R691A) SpCas9 and 16 µg MS sgRNA (pre-complexed for 10 min at room temperature; HiFi RNP). HSPCs were resuspended with HiFi RNP in P3 buffer and electroporated using program DZ-100 on the Lonza 4D nucleofector. Immediately following electroporation, $CD34^+$ HSPCs were transduced with HBB-specific AAV6 barcoded donor template libraries at 2500–5000 vector genomes per cell and 20000 vector genomes per cell for AAVS1-specific AAV6 barcoded libraries. 12-16 h post-transduction, targeted cells were washed and resuspended in fresh media and allowed to culture for additional 24-36 h, with a total manufacturing time less than 96 h.

In vitro erythrocyte differentiation of HBB-targeted $CD34^+$ HSPCs. SCD-HSPCs were targeted with either the therapeutic AAV6 donor (with one sequence) or the HBB barcoded AAV6 donor template library and subjected to the in vitro erythrocyte differentiation protocol two days post targeting as previously described[32,62,63]. Base medium was supplemented with 100 U/mL of penicillin–streptomycin, 10 ng/mL SCF, 1 ng/mL IL-3 (PeproTech), 3 U/mL erythropoietin (eBiosciences), 200 µg/mL transferrin (Sigma-Aldrich), 3% antibody serum (heat-inactivated from Atlanta Biologicals, Flowery Branch, GA, USA), 2% human plasma (umbilical cord blood), 10 µg/mL insulin (Sigma Aldrich) and 3 U/mL heparin (Sigma-Aldrich). Briefly, targeted HSPCs were differentiated into erythrocytes using a three-phase differentiation protocol that lasted 14–16 days in culture. Cultures were maintained at a concentration of 2.5e5/mL. The first phase of erythroid differentiation corresponded to days 0–7 (day 0 being day 2 after electroporation). During the second phase of differentiation, corresponding to days 7-10, IL-3 was discontinued from the culture medium. In the third and final phase, corresponding to days 10–16, transferrin was increased to 1 mg/mL. Differentiated cells were then harvested for analysis of hemoglobin tetramers by cation-exchange high performance liquid chromatography.

Hemoglobin tetramer analysis via cation-exchange HPLC. Hemoglobin tetramer analysis was performed as previously described[32]. Briefly, red blood cell pellets were flash frozen post differentiation until tetramer analysis when pellets were then thawed, lysed with 3 times volume of water, incubated for 15 min, and then sonicated for 30 s to finalize the lysing procedure. Cells were then centrifuged for 5 min at 13,000 rpm and used for input to analyze steady-state hemoglobin tetramer levels. Transfused blood from sickle cell disease patients was always used to ascertain the retention time of sickle, adult, and fetal human hemoglobin.

Transplantation of targeted $CD34^+$ HSPCs into NSG mice. Six to eight-week-old immunodeficient NSG female mice were sublethally irradiated with 200 cGy 12–24 h before injection of cells. For primary transplants, 2–4 × 10^5 targeted $CD34^+$ HSPCs were harvested two days post electroporation, spun down at 300 g, and resuspended in 25 µl PBS before intrafemoral (IF) transplantation into the right femur of female NSG mice. A total of 4 mice (two HBB and two BFP) were selected for secondary transplants based on high genome editing and xenograft efficiencies. Cells from 3 of these mice were successfully transplanted into secondary recipients. One mouse died due to complications associated with intrafemoral injection. Mononuclear cells (MNCs) were harvested from primary transplanted NSG mice, and half of the total MNCs were used to transplant one sublethally irradiated female NSG mouse via IF injection.

Analysis of human engraftment and fluorescent activated cell sorting. In total 16–18 weeks following transplantation of targeted HSPCs, mice were euthanized, bones (2x femurs, 2x pelvis, 2x tibia, sternum, spine) were collected and crushed as previously described[36,60]. MNCs were harvested by ficoll gradient centrifugation and human hematopoietic cells were identified by flow cytometry using the following antibody cocktail: HLA-A/B/C FITC (1:100, clone W6/32, Biolegend), mouse CD45.1 PE-CY7 (1:200, clone A20, Thermo Scientific), CD34 APC (1:100, clone 581, Biolegend), CD33 V450 (1:100, clone WM53, BD Biosciences), CD19 Percp5.5 (1:50, clone HIB19, BD Biosciences), CD10 APC-Cy7 (1:20, HI10a, Biolegend), mTer119 PeCy5 (1:200, clone Ter-119, Thermo Scientific), and CD235a PE (1:50, HIR2, Thermo Scientific). For mice transplanted with AAVS1-edited HSPCs, the following cocktail was used: HLA-A/B/C FITC (1:100, clone W6/32, Biolegend), mouse CD45.1 PE-CY7 (1:200, clone A20, Thermo Scientific), CD34 APC (1:50, clone 581, Biolegend), CD33 PE (1:100, clone WM53, BD Biosciences), CD19 BB700 (1:50, clone HIB19, BD Biosciences), CD3 APC-Cy7

(1:50, clone SK7, BD Biosciences). For AAVS1-edited HSPCs, CD33^{Hi} and CD33^{Mid} were sorted individually, however the data were aggregated for analysis. Human hematopoietic cells were identified as HLA-A/B/C positive and mCD45.1 negative. The following gating scheme was used to sort cell lineages to be analyzed for barcoded recombination alleles: Myeloid cells (CD33^{+}), B Cells (CD19^{+}), HSPCs (CD10^{-}, CD34^{+}, CD19$^{−}$, CD33$^{−}$), and erythrocytes (Ter119$^{−}$, mCD45.1^{-}, CD19$^{−}$, CD33-, CD10$^{−}$, CD235a^{+}). Sorted cells were spun down, genomic DNA was harvested using QuickExtract (Lucigen), and was saved until library preparation and sequencing.

Sequencing library preparation. Harvested cells were lysed using QuickExtract DNA Extraction Solution (Lucigen, Cat. No. QE09050) following manufacturers protocol. Based on the starting cell count, 0.5–1 μL QuickExtract lysate was used for PCR. Spike-in experiments were designed to simulate the amplification of known alleles by spiking 0.1 ng plasmid into 60 ng gDNA. All PCRs for library preparation was carried out using Q5 High-Fidelity 2X Master Mix (NEB, Cat. No. M0492L). An initial enrichment amplification of 15 cycles was followed with a second round of PCR using unique P5 and P7 indexing primer combinations for 15 cycles and purified using 1.8X SPRI beads. For nested PCR (used to specifically amplify gDNA from in vitro experiments, in which unintegrated rAAV donor DNA may confound results), an initial amplification of 30 cycles was used. PCR products were analyzed by gel electrophoresis and purified using 1X SPRI beads.

PCR products were normalized, pooled, and then gel extracted using the QIAEX II Gel Extraction Kit (Qiagen, Cat. No. 20051). The resulting libraries were sequenced using both Illumina Miseq (2 × 150 bp paired-end) and Illumina HiSeq 4000 (2 × 150 bp paired-end) platforms. Illumina HiSeq 4000 sequencing was performed by Novogene Corporation.

Index switching correction of false-positive NGS reads. We utilized two independent methods to determine the incidence of index-switching present in samples that were run on a HiSeq 4000[64,65]. In one approach, we calculated the number of contaminating reads between two different amplicons sequenced in the same pool. As a second approach, we utilized the algorithm developed by Larrson et al. to estimate the fraction of reads which were spread to other samples through index switching[66]. Both of these methods yielded an index switching incidence of 0.3%. We performed a conservative correction for this by subtracting 0.3% x [# Barcode Reads] from each barcode in each sample (as described in Supplementary Fig 1).

Statistical analysis. All statistical tests used in this study were performed using GraphPad Prism 7/8 or R version 3.6.1. For comparing the average of two means, we used the Student's t-test to reject the null hypothesis ($P < 0.05$).

Informed consent and use of laboratory animals. Patients with SCD consented to the use of CD34^{+} HSPCs for research with the accompanying IRB approval. All mouse experiments were conducted in accordance with a protocol approved by the Institutional Animal Care and Use Committee (Stanford Administrative Panel on Laboratory Animal Care no. 22264).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The authors declare that all data supporting the findings of this study are available within the article and its supplementary information files or from the corresponding author upon reasonable request.

Barcode counts after processing by the TRACEseq pipeline have been deposited in the GEO database under accession code: GSE155910.

Code availability

Code used to generate bubble plots are available on github (https://github.com/armonazizi/TRACEseq).

Received: 6 April 2020; Accepted: 16 December 2020;
Published online: 20 January 2021

References

1. High, K. A. & Roncarolo, M. G. Gene therapy. *N. Engl. J. Med* **381**, 455–464 (2019).
2. Cavazzana, M., Bushman, F. D., Miccio, A., Andre-Schmutz, I. & Six, E. Gene therapy targeting haematopoietic stem cells for inherited diseases: progress and challenges. *Nat. Rev. Drug Disco.* **18**, 447–462 (2019).
3. Biasco, L. et al. In vivo tracking of T cells in humans unveils decade-long survival and activity of genetically modified T memory stem cells. *Sci. Transl. Med.* **7**, 273ra213 (2015).
4. Biasco, L. et al. In vivo tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases. *Cell Stem Cell* **19**, 107–119 (2016).
5. Scala, S. et al. Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans. *Nat. Med.* **24**, 1683–1690 (2018).
6. Mamcarz, E. et al. Lentiviral gene therapy combined with low-dose busulfan in infants with SCID-X1. *N. Engl. J. Med.* **380**, 1525–1534 (2019).
7. Fraietta, J. A. et al. Disruption of TET2 promotes the therapeutic efficacy of CD19$^{−}$targeted T cells. *Nature* **558**, 307–312 (2018).
8. Marktel, S. et al. Intrabone hematopoietic stem cell gene therapy for adult and pediatric patients affected by transfusion-dependent ss-thalassemia. *Nat. Med.* **25**, 234–241 (2019).
9. Porter, S. N., Baker, L. C., Mittelman, D. & Porteus, M. H. Lentiviral and targeted cellular barcoding reveals ongoing clonal dynamics of cell lines in vitro and in vivo. *Genome Biol.* **15**, R75 (2014).
10. Lu, R., Neff, N. F., Quake, S. R. & Weissman, I. L. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* **29**, 928–933 (2011).
11. Yabe, I. M. et al. Barcoding of macaque hematopoietic stem and progenitor cells: a robust platform to assess vector genotoxicity. *Mol. Ther. Methods Clin. Dev.* **11**, 143–154 (2018).
12. Wu, C. et al. Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. *Cell Stem Cell* **14**, 486–499 (2014).
13. Kristiansen, T. A. et al. Cellular barcoding links B-1a B cell potential to a fetal hematopoietic stem cell state at the single-cell level. *Immunity* **45**, 346–357 (2016).
14. Thielecke, L. et al. Limitations and challenges of genetic barcode quantification. *Sci. Rep.* **7**, 43249 (2017).
15. Barzel, A. et al. Promoterless gene targeting without nucleases ameliorates haemophilia B in mice. *Nature* **517**, 360–364 (2015).
16. Russell, D. W. & Hirata, R. K. Human gene targeting by viral vectors. *Nat. Genet.* **18**, 325–330 (1998).
17. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
18. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* (2019). https://doi.org/10.1038/s41586-019-1711-4
19. Miller, D. G., Petek, L. M. & Russell, D. W. Human gene targeting by adeno-associated virus vectors is enhanced by DNA double-strand breaks. *Mol. Cell Biol.* **23**, 3550–3557 (2003).
20. Porteus, M. H. & Baltimore, D. Chimeric nucleases stimulate gene targeting in human cells. *Science* **300**, 763 (2003).
21. Genovese, P. et al. Targeted genome editing in human repopulating haematopoietic stem cells. *Nature* **510**, 235–240 (2014).
22. Urnov, F. D. et al. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**, 646–651 (2005).
23. Porteus, M. H. & Carroll, D. Gene targeting using zinc finger nucleases. *Nat. Biotechnol.* **23**, 967–973 (2005).
24. Lombardo, A. et al. Site-specific integration and tailoring of cassette design for sustainable gene transfer. *Nat. Methods* **8**, 861–869 (2011).
25. Romero, Z. et al. Editing the sickle cell disease mutation in human hematopoietic stem cells: comparison of endonucleases and homologous donor templates. *Mol. Ther.* **27**, 1389–1406 (2019).
26. Pattabhi, S. et al. In vivo outcome of homology-directed repair at the HBB gene in HSC using alternative donor template delivery methods. *Mol. Ther. Nucleic Acids* **17**, 277–288 (2019).
27. Magis, W. et al. High-level correction of the sickle mutation amplified in vivo during erythroid differentiation. *bioRxiv*, 432716 (2019).
28. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
29. Qi, L. S. et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).
30. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
31. Canver, M. C. et al. Characterization of genomic deletion efficiency mediated by clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells. *J. Biol. Chem.* **289**, 21312–21324 (2014).
32. Vakulskas, C. A. et al. A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nat. Med.* **24**, 1216–1224 (2018).
33. Porteus, M. H. A new class of medicines through DNA editing. *N. Engl. J. Med.* **380**, 947–959 (2019).

34. Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J. S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res* **24**, 1012–1019 (2014).
35. Martin, R. M. et al. Highly efficient and marker-free genome editing of human pluripotent stem cells by CRISPR-Cas9 RNP and AAV6 donor-mediated homologous recombination. *Cell Stem Cell* **24**, 821–828 e825 (2019).
36. Dever, D. P. et al. CRISPR/Cas9 beta-globin gene targeting in human haematopoietic stem cells. *Nature* **539**, 384–389 (2016).
37. Pavel-Dinu, M. et al. Gene correction for SCID-X1 in long-term hematopoietic stem cells. *Nat. Commun.* **10**, 1634 (2019).
38. Schiroli, G. et al. Preclinical modeling highlights the therapeutic potential of hematopoietic stem cell gene editing for correction of SCID-X1. *Sci. Transl. Med.* **9**, eaan0820 (2017).
39. Gomez-Ospina, N. et al. Human genome-edited hematopoietic stem cells phenotypically correct Mucopolysaccharidosis type I. *Nat. Commun.* **10**, 4045 (2019).
40. De Ravin, S. S. et al. CRISPR-Cas9 gene repair of hematopoietic stem cells from patients with X-linked chronic granulomatous disease. *Sci. Transl. Med.* **9**, eaah3480(2017).
41. Hubbard, N. et al. Targeted gene editing restores regulated CD40L function in X-linked hyper-IgM syndrome. *Blood* **127**, 2513–2522 (2016).
42. Eyquem, J. et al. Targeting a CAR to the TRAC locus with CRISPR/Cas9 enhances tumour rejection. *Nature* **543**, 113–117 (2017).
43. van Overbeek, M. et al. DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks. *Mol. Cell* **63**, 633–646 (2016).
44. Rogers, Z. N. et al. A quantitative and multiplexed approach to uncover the fitness landscape of tumor suppression in vivo. *Nat. Methods* **14**, 737–742 (2017).
45. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
46. Hendel, A. et al. Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nat. Biotechnol.* **33**, 985–989 (2015).
47. Doulatov, S., Notta, F., Laurenti, E. & Dick, J. E. Hematopoiesis: a human perspective. *Cell Stem Cell* **10**, 120–136 (2012).
48. Davidsson, M. et al. A novel process of viral vector barcoding and library preparation enables high-diversity library generation and recombination-free paired-end sequencing. *Sci. Rep.* **6**, 37563 (2016).
49. Cheung, A. M. et al. Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice. *Blood* **122**, 3129–3137 (2013).
50. Reinisch, A. et al. A humanized bone marrow ossicle xenotransplantation model enables improved engraftment of healthy and leukemic human hematopoietic cells. *Nat. Med* **22**, 812–821 (2016).
51. Rongvaux, A. et al. Corrigendum: development and function of human innate immune cells in a humanized mouse model. *Nat. Biotechnol.* **35**, 1211 (2017).
52. Wunderlich, M. et al. AML xenograft efficiency is significantly improved in NOD/SCID-IL2RG mice constitutively expressing human SCF, GM-CSF and IL-3. *Leukemia* **24**, 1785–1788 (2010).
53. Notta, F. et al. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science* **351**, aab2116 (2016).
54. Canny, M. D. et al. Inhibition of 53BP1 favors homology-dependent DNA repair and increases CRISPR-Cas9 genome-editing efficiency. *Nat. Biotechnol.* **36**, 95–102 (2018).
55. Schiroli, G. et al. Precise gene editing preserves hematopoietic stem cell function following transient p53-Mediated DNA damage response. *Cell Stem Cell* **24**, 551–565 e558 (2019).
56. Wilkinson, A. C. et al. Author Correction: Long-term ex vivo haematopoietic-stem-cell expansion allows nonconditioned transplantation. *Nature* **571**, E12 (2019).
57. Fares, I. et al. Cord blood expansion. Pyrimidoindole derivatives are agonists of human hematopoietic stem cell self-renewal. *Science* **345**, 1509–1512 (2014).
58. Cohen, S. et al. Hematopoietic stem cell transplantation using single UM171-expanded cord blood: a single-arm, phase 1-2 safety and feasibility study. *Lancet Haematol.* **7**, e134–e145 (2020).
59. Bai, T. et al. Expansion of primitive human hematopoietic stem cells by culture in a zwitterionic hydrogel. *Nat. Med.* **25**, 1566–1575 (2019).
60. Bak, R. O., Dever, D. P. & Porteus, M. H. CRISPR/Cas9 genome editing in human hematopoietic stem cells. *Nat. Protoc.* **13**, 358–376 (2018).
61. Aurnhammer, C. et al. Universal real-time PCR for the detection and quantification of adeno-associated virus serotype 2-derived inverted terminal repeat sequences. *Hum. Gene Ther. Methods* **23**, 18–28 (2012).
62. Dulmovits, B. M. et al. Pomalidomide reverses gamma-globin silencing through the transcriptional reprogramming of adult hematopoietic progenitors. *Blood* **127**, 1481–1492 (2016).
63. Hu, J. et al. Isolation and functional characterization of human erythroblasts at distinct stages: implications for understanding of normal and disordered erythropoiesis in vivo. *Blood* **121**, 3246–3253 (2013).
64. Costello, M. et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **19**, 332 (2018).
65. Sinha, R. et al. Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *bioRxiv*, 125724 (2017).
66. Larsson, A. J. M., Stanley, G., Sinha, R., Weissman, I. L. & Sandberg, R. Computational correction of index switching in multiplexed sequencing libraries. *Nat. Methods* **15**, 305–307 (2018).

Acknowledgements

R.S. was supported by a Postdoctoral Fellowship, PF-18-102-01-DOC, from the American Cancer Society. R.M. acknowledges support for the Stanford Ludwig Center for Cancer Stem Cell Research and Medicine, and NIH grants R01-CA188055 and R01-HL142637. M.H.P. gratefully acknowledges the support of the Amon Carter Foundation, the Laurie Kraus Lacob Faculty Scholar Award in Pediatric Translational Research and NIH grant support R01-AI097320 and R01-AI120766. G.B. acknowledges support from the Cancer Prevention and Research Institute of Texas (RR14008 and RP170721). D.P.D and J.C. were supported by NIH grant support R01-HL135607. We thank the Binns Program for Cord Blood Research at Stanford University for cord-blood-derived CD34$^+$ HSPCs and also for SCD-HSPCs.

Author contributions

R.S., D.P.D., R.M., and M.H.P. conceived the study. R.S. and D.P.D. designed and performed experiments. A.A. designed and performed bioinformatic analyses. C.M.L., and Y.P. performed the library preparation for MiSeq and HiSeq runs with support from G.B. J.C. and T.K. developed reagents and performed analyses. R.S., D.P.D., A.A., R.M., and M.H.P. wrote the manuscript with support from all authors. R.S. created figures with support from all authors.

Competing interests

R.S. and D.P.D. are presently employed by Graphite Bio. M.H.P. has equity and serves on the scientific advisory board of CRISPR Therapeutics and Allogene Therapeutics. R.M. is on the Board of Directors of BeyondSpring Inc, the Scientific Advisory Boards of Zenshine Pharmaceuticals and Kodikaz Therapeutic Solutions Inc., and is an inventor on a number of patents related to CD47 cancer immunotherapy licensed to Gilead Sciences, Inc. However, none of these companies had input into the design, execution, interpretation, or publication of the work in this manuscript. Other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-20792-y.

Correspondence and requests for materials should be addressed to G.B., M.H.P. or R.M.

Peer review information *Nature Communications* thanks Rong Lu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints