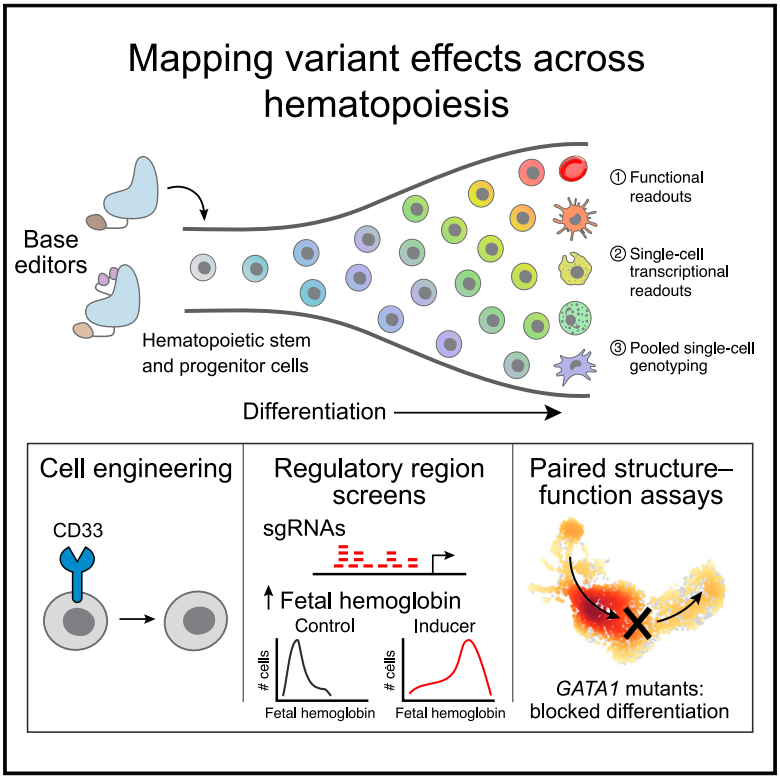
Cell

Massively parallel base editing to map variant effects in human hematopoiesis

Graphical abstract



Highlights

- Base-editing screens map variant effects in relevant hematopoietic cell types
- Hematopoietic stem cell screens could enable improved anti-CD33 leukemia therapy
- Screens identify non-coding variants that modulate fetal hemoglobin expression
- Single-cell readouts dissect variant pathogenicity and impact on differentiation

Authors

Jorge D. Martin-Rufino, Nicole Castano, Michael Pang, ..., Eric S. Lander, Daryl E. Klein, Vijay G. Sankaran

Correspondence

sankaran@broadinstitute.org

In brief

Genomic screening of hematopoietic stem and progenitor cells is accomplished at large scale by combining base editor-mediated gene perturbation with functional and sequencing readouts, and it enables interrogation of diverse aspects of blood cell biology and cellular engineering.

Martin-Rufino et al., 2023, Cell *186*, 2456–2474
May 25, 2023
https://doi.org/10.1016/j.cell.2023.03.035

CellPress

CellPress
OPEN ACCESS

Cell

Resource

Massively parallel base editing to map variant effects in human hematopoiesis

Jorge D. Martin-Rufino,[1,2,3] Nicole Castano,[1,2,18] Michael Pang,[1,2,4,18] Emanuelle I. Grody,[2,17,18] Samantha Joubran,[1,2,5] Alexis Caulier,[1,2] Lara Wahlster,[1,2] Tongqing Li,[6] Xiaojie Qiu,[2,7] Anna Maria Riera-Escandell,[8] Gregory A. Newby,[2,9,10] Aziz Al'Khafaji,[2] Santosh Chaudhary,[2] Susan Black,[1,2] Chen Weng,[1,2,7] Glen Munson,[2] David R. Liu,[2,9,10] Marcin W. Wlodarski,[11] Kacie Sims,[12] Jamie H. Oakley,[13] Ross M. Fasano,[13,16] Ramnik J. Xavier,[2,14] Eric S. Lander,[2] Daryl E. Klein,[6] and Vijay G. Sankaran[1,2,15,19,*]

[1]Division of Hematology/Oncology, Boston Children's Hospital and Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA
[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[3]PhD Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115, USA
[4]Harvard-MIT Health Sciences and Technology, Harvard Medical School, Boston, MA 02115, USA
[5]Chemical Biology PhD Program, Harvard Medical School, Boston, MA 02115, USA
[6]Department of Pharmacology and Yale Cancer Biology Institute, Yale University School of Medicine, New Haven, CT 06510, USA
[7]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA
[8]Independent Researcher, now at Google, Zurich, Switzerland
[9]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA
[10]Howard Hughes Medical Institute, Harvard University, Cambridge, MA 02138, USA
[11]Department of Hematology, St Jude Children's Research Hospital, Memphis, TN 38105, USA
[12]St. Jude Affiliate Clinic at Our Lady of the Lake Children's Health, Baton Rouge, LA 70809, USA
[13]Aflac Cancer and Blood Disorders Center, Children's Healthcare of Atlanta and Emory University, Atlanta, GA 30322, USA
[14]Center for Computational and Integrative Biology, Department of Molecular Biology, and Center for the Study of Inflammatory Bowel Disease, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA
[15]Harvard Stem Cell Institute, Cambridge, MA 02138, USA
[16]Center for Transfusion Medicine and Cellular Therapies, Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA 30322, USA
[17]Present address: Driskill Graduate Program in Life Sciences, Northwestern University, Chicago, IL 60611, USA
[18]These authors contributed equally
[19]Lead contact
*Correspondence: sankaran@broadinstitute.org
https://doi.org/10.1016/j.cell.2023.03.035

SUMMARY

Systematic evaluation of the impact of genetic variants is critical for the study and treatment of human physiology and disease. While specific mutations can be introduced by genome engineering, we still lack scalable approaches that are applicable to the important setting of primary cells, such as blood and immune cells. Here, we describe the development of massively parallel base-editing screens in human hematopoietic stem and progenitor cells. Such approaches enable functional screens for variant effects across any hematopoietic differentiation state. Moreover, they allow for rich phenotyping through single-cell RNA sequencing readouts and separately for characterization of editing outcomes through pooled single-cell genotyping. We efficiently design improved leukemia immunotherapy approaches, comprehensively identify non-coding variants modulating fetal hemoglobin expression, define mechanisms regulating hematopoietic differentiation, and probe the pathogenicity of uncharacterized disease-associated variants. These strategies will advance effective and high-throughput variant-to-function mapping in human hematopoiesis to identify the causes of diverse diseases.

INTRODUCTION

An urgent need for the study of human physiology and disease is the ability to efficiently introduce large numbers of specific single-base substitutions in endogenous loci in primary human cells.[1] This ability would enable, in the context of natural regulation in disease-relevant cell types, a wide variety of applications, including systematic studies of the roles of amino acids across coding regions, the consequences of mutations near splice sites, the function of non-coding genetic variants identified by genome-wide association studies, the architecture of enhancers, and the design of gene therapy strategies.

Existing scalable approaches provide valuable information but have major limitations. Massively parallel reporter assays on

plasmids, for example, are largely confined to cell lines and do not reflect endogenous regulation. CRISPR-Cas9 screens involving cutting, inhibition, and activation can alter the overall expression of a gene, but they cannot interrogate variants at single-base resolution. Moreover, CRISPR-Cas9 cutting causes heterogeneous collections of small and large insertions and deletions from repair of DNA double-stranded breaks, as well as complex chromosomal rearrangements and selection of cells with p53 suppression.[2–6]

Base editors present a valuable alternative to these approaches because they enable the creation of specific single-nucleotide changes, which are the most common type of genetic variation present in the human genome and the cause of most genetic disease. Specifically, adenine base editors (ABEs) produce A·T to G·C changes that can correct ~50% of pathogenic point mutations observed in humans, and cytosine base editors (CBEs) produce C·G to T·A changes that can correct ~14% of pathogenic point mutations in humans.[7–11] The recent development of C·G to G·C and A·T to T·A base editors now enables the creation of nearly all types of point mutations.[12–16]

Recent studies have demonstrated the ability to use base editors to conduct systematic screens in cell lines.[17–21] As we learn more from studies of human cell atlases, it is clear that the diversity of cell states present in human physiology and disease is strikingly varied, and disease-associated variants often have impacts on specific cellular states and circuits, which may be poorly represented by existing cell lines.[22,23] There is a pressing need to be able to perform screens in primary cells.

One major challenge is the inability to readily modify the genome in primary cells in order to express Cas9 derivatives in a stable manner to allow for efficient editing.[24,25] Moreover, many cell states are transient and only rarely observed. An important example is the cells observed in human hematopoiesis, that is, the production of blood and immune cells.[26] The ability to perform large-scale endogenous single-nucleotide perturbations in primary cells is critical to understanding the large number of variants that are associated with a spectrum of blood and immune cell disorders, as well as variation in hematopoiesis.[26–33]

A second major challenge is the need to assess cells across a wide variety of cell states in primary tissues, such as in hematopoiesis. This requires general and sensitive readouts, as can be achieved by using single-cell RNA sequencing (scRNA-seq).

Here, we develop and use base editor screens in primary hematopoietic stem and progenitor cells (HSPCs), both in the undifferentiated state or upon directed differentiation, to provide key insights into and address a broad set of biological problems (Figure 1A). Not only do we conduct functional screens in primary human HSPCs, but we also use scRNA-seq readouts, as has been done for Perturb-seq using other genome-editing approaches (Figure 1B).[34–37] Moreover, we demonstrate the utility of pooled single-cell genotyping performed separately to rapidly and efficiently assess editing outcomes and facilitate screen interpretation.

The development of high-throughput assays that link variants to complex cellular phenotypes in primary human hematopoietic cells opens the door to functionally interrogate an entire tissue at single-variant resolution, and systematically decode the effect of such genetic variation on a diverse group of hematologic, immunologic, oncologic, metabolic, neurologic, and inflammatory diseases that result from alterations in blood and immune cells, as well as their precursors.

RESULTS

Achieving highly efficient base editing in primary hematopoietic cells with detectable lentiviral readouts

To enable massively parallel variant screens in primary human hematopoietic cells, we needed to deliver both base editors and pooled single-guide RNAs (sgRNAs). While screens in cell lines can be performed by establishing stable expression of Cas9 derivatives such as base editors, stable lentiviral transduction of these genome-editing tools is not feasible in primary HSPCs.[24,25,38] As an initial approach, we electroporated recombinant base editor protein precomplexed with sgRNA guides. Specifically, we used the latest generation of ABEs (ABE8e), and used a Cas9 that recognizes a relaxed motif (NG) as its protospacer adjacent motif (PAM), in order to expand the number of targetable sites (Figure S1A; STAR Methods).[39] This approach achieved editing efficiencies of ~80% of alleles in HSPCs for four genomic targets tested, while maintaining cell viability (Figures 1C and S1B; STAR Methods).

We next adapted the protocol to enable massively parallel screens. We infected primary HSPCs with lentiviruses expressing sgRNAs to make it possible to identify the sgRNA present in each cell, as done in other types of Cas9 screens,[40,41] and 2 days later electroporated ABE8e protein. We observed a dose-dependent increase in editing efficiency of up to 65% as a function of electroporated protein concentration using this approach, while allowing for detection of the sgRNA identity in each cell (Figures 1D and S1C; STAR Methods). Although prior reports have suggested that precomplexing Cas9 protein with non-targeting (NT) sgRNAs can increase editing from lentivirally encoded sgRNAs,[41] we observed better editing with delivery of the base editor protein alone (Figure S1D).[40] Finally, we also employed highly efficient library assembly strategies (Golden Gate, STAR Methods) for all screens we conducted with this approach (Figures S1E–S1G; STAR Methods).

Functional base-editing screens in human HSCs to improve cell therapy

As a first test, we applied our method to an application in cancer immunotherapy. Advances in CRISPR-Cas9 genome editing of human HSPCs have enabled therapeutic strategies for diverse diseases.[42] Although some of these approaches have entered the clinic, concerns have been raised about undesirable impacts of the double-strand DNA breaks produced by CRISPR-Cas9.[2–6] A more controlled and potentially safer approach to abrogate gene function might be to use base editors to alter key splice sites in a target gene.[43]

To test whether we could perform functional screens using base editors in primary HSPCs, we performed a systematic mutagenesis screen of splice sites in CD33. CD33 is a key target in immunotherapy against acute myeloid leukemia (AML) because it is expressed on AML cells.[44–46] Unfortunately, CD33 is also expressed on normal hematopoietic cells, including hematopoietic stem cells (HSCs), which limits the clinical

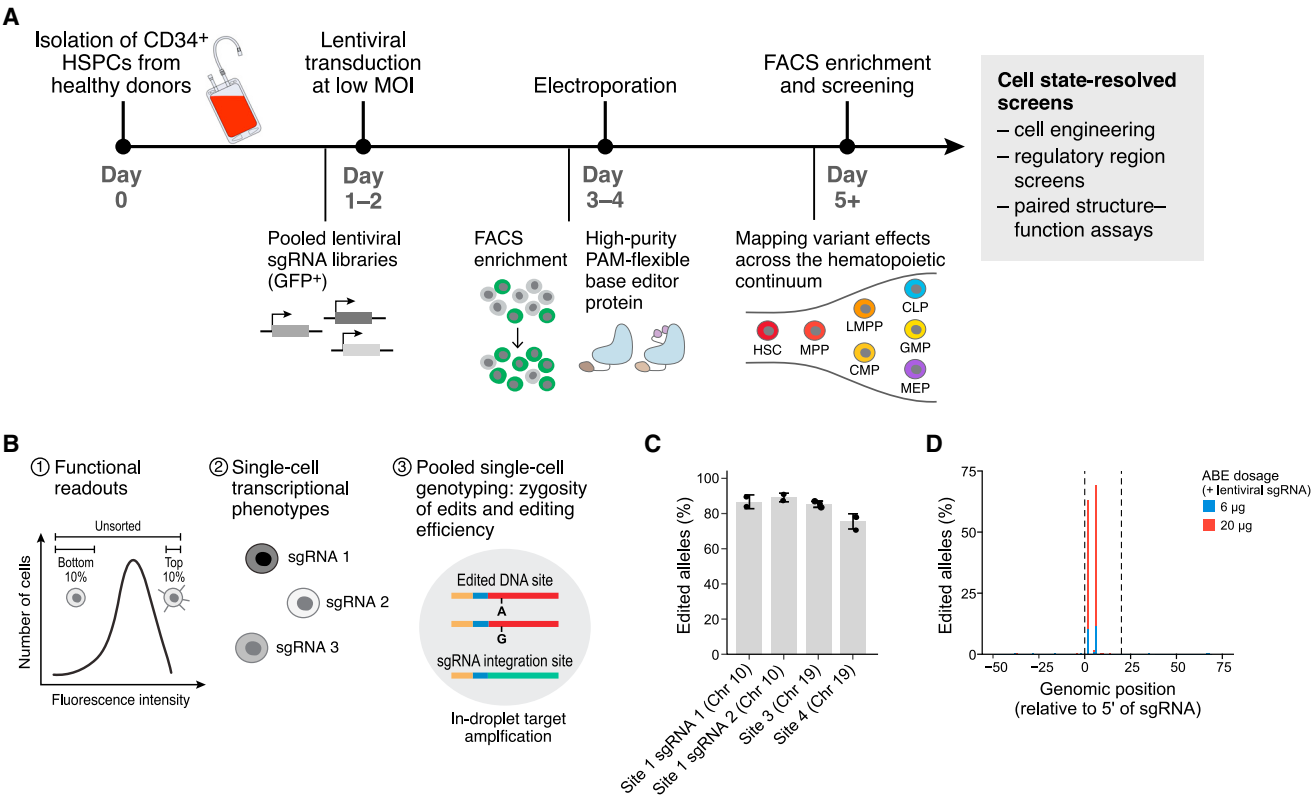


Figure 1. Massively parallel variant assessment in primary hematopoiesis enabled by purified base editor protein delivery and lentiviral sgRNA transduction

(A) Schematic of base editor screens in hematopoiesis.

(B) Schematic of the readouts used to analyze variant effects.

(C) Editing efficiencies using purified ABE8e-Cas9NG protein and chemically modified sgRNAs across four genomic targets (n = 2–3 independent electroporations per site), using 20 µg of ABE8e.

(D) Editing efficiencies using lentivirally transduced sgRNAs on site 3 (chromosome 19), as a function of ABE protein dosage in HSPCs.

See also Figure S1.

effectiveness of CD33 as a target for CAR-T cell immunotherapy (Figure S2A). It has recently been shown that this limitation can be overcome by infusing patients with human HSCs in which CD33 has been knocked out by Cas9-mediated genome editing, making it possible to eradicate AML cells by anti-CD33 CAR-T therapy, while preserving hematopoiesis.[44–46]

We performed a base-editing screen against all canonical splice donor or acceptor sites present in *CD33* to identify single-base edits that reduced or eliminated CD33 expression in HSPCs (Figure 2A). We used fluorescence-activated cell sorting (FACS) to select the 10% of cells with the lowest CD33 levels ($CD33^{low}$) and compared the sgRNAs found in this population with those in the unsorted population ($CD33^{unsorted}$). Multiple sgRNAs were enriched in the $CD33^{low}$ population, with strong enrichment for sgRNAs targeting splice donor sites of exons 1 and 2 (Figure 2B). Negative control sgRNAs that cause silent mutations or are NT showed no enrichment. A second biological replicate (with independent library cloning and HSPCs from a different human donor) displayed high concordance (Figure S2B).

We individually validated the top candidates, the splice donor sgRNAs against exons 1 and 2, by using ribonucleoproteins of ABE precomplexed with chemically modified sgRNAs, and we compared the results with those produced by double-stranded cutting by (nuclease active) Cas9.[44–46] The base-edited cells exhibited near-complete absence of CD33 expression as assessed by flow cytometry, comparable to the results of Cas9-mediated disruption (Figure 2C), and by measurement of mRNA levels[44,45] (Figure S2D). While both ABE and Cas9 achieved equivalent editing efficiencies (>90%, Figure S2C), the base-edited cells showed a highly homogeneous pattern of edits, while the Cas9-edited cells showed varied insertion-deletion (indel) patterns (Figure S2E).

Having confirmed editing in a bulk population of HSPCs, we next showed that the most primitive HSPC compartment necessary for long-term hematopoietic maintenance,[42,47] marked by CD34+CD90+CD45RA− surface expression displayed near-complete elimination of CD33 expression (Figure 2D). To functionally confirm the long-term repopulating potential of CD33-base-edited HSPCs, we transplanted cells targeted with exon 1 splice donor or NT control sgRNAs into immunodeficient and Kit mutant mouse recipients (NOD.Cg-*Kit*$^{W\text{-}41J}$*Tyr*$^{+}$*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/ThomJ, NBSGW, Figure 2E).[48] Human CD33 base-edited cells displayed comparable engraftment to control cells

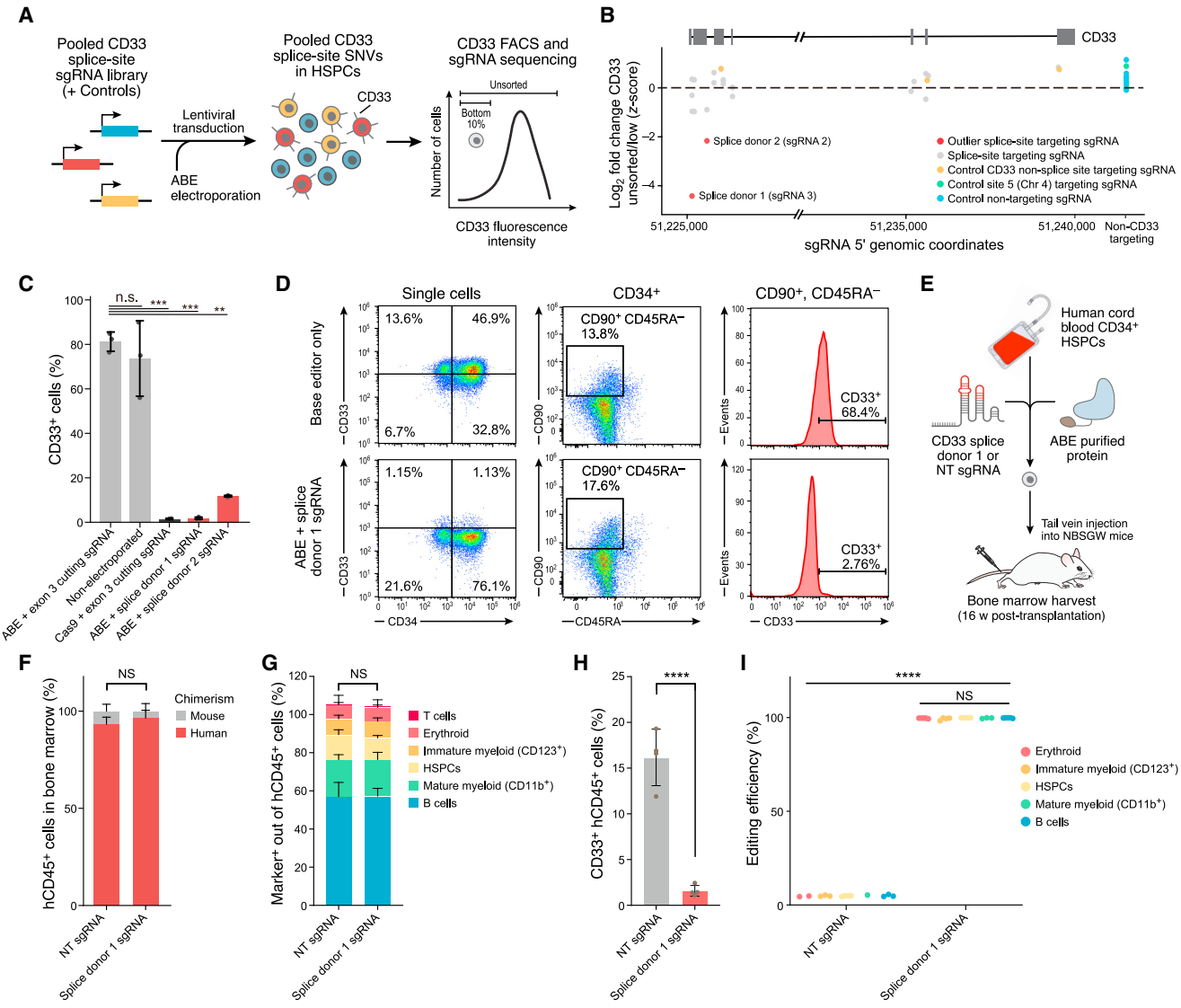


Figure 2. Splice-site base editor screens in primary hematopoietic stem and progenitor cells for improved cell therapies

(A) Schematic of the screen design using adenine base editor ABE8e-Cas9NG targeting all CD33 splice sites in HSPCs.

(B) *Z* scored $\log_2$(FC) in sgRNA reads between HSPCs with the bottom 10% CD33 levels and the unsorted population. Plotted are the 5′ genomic coordinates of each sgRNA. CD33 non-splice-site-targeting sgRNAs, non-targeting sgRNAs and an sgRNA targeting site 5 (chromosome 4) (Figure S1C) are shown as controls.

(C) Percentage of CD33 expressing cells, as assessed by flow cytometry (n = 3 independent electroporations). Two-tailed unpaired t test.

(D) Flow-cytometry comparison of CD34+, CD90+, CD45RA− HSPCs in splice donor 1 base-edited and control cells, and percentage of CD33-expressing cells in that population. Representative data from three independent electroporations.

(E) Schematic of the experiment to assess the *in vivo* engraftment potential of cord blood-derived HSPCs electroporated with CD33 splice donor 1 or non-targeting ABE ribonucleoproteins into NBSGW mice.

(F) Engraftment of CD33 splice donor 1 base-edited HSPCs in immunocompromised mice at 16 weeks post-transplantation. Percentage of human CD45+ cells in mouse bone marrow for mice transplanted with CD33 splice donor 1 or non-targeting ABE ribonucleoproteins (n = 5 mice for CD33 splice donor 1 and n = 4 mice for non-targeting). Two-tailed unpaired t test for each population.

(G) Percentages of the main human hematopoietic lineages measured in mouse bone marrow for mice transplanted with HPSCs edited with CD33 splice donor 1 or non-targeting ABE ribonucleoproteins (n = 5 mice for CD33 splice donor 1 and n = 4 mice for non-targeting). Two-tailed unpaired t test for each lineage.

(H) CD33 base-editing-mediated KO in human CD45+ cells in the mouse bone marrow assessed by FACS (n = 4 mice for CD33 splice donor 1 and n = 4 mice for non-targeting). Two-tailed unpaired t test.

(I) Percentage of edited reads for each of the FACS-purified human lineages from bone marrow for mice transplanted with CD33 splice donor 1 or non-targeting ABE ribonucleoproteins. For each lineage, each dot represents samples from a mouse. Low quality amplicon samples were excluded. Two-way ANOVA.

See also Figure S2.

with similar lineage composition in the bone marrow, as assayed after long-term hematopoietic reconstitution was achieved 16 weeks after transplantation (Figures 2F, 2G, and S2F–S2H). Transplanted human cells edited with the exon 1 splice donor sgRNA demonstrated a marked reduction in CD33 expression, both overall (Figures 2H and S2G) and in the HSPC compartment (Figures S2F and S2G). The editing efficiency observed in the FACS-purified erythroid, myeloid, B cell, and HSPC compartments was consistently greater than 95% (Figure 2I). Our results demonstrate the utility of a functional screen in primary hematopoietic cells to identify effective alternative cell therapy approaches that warrant further investigation in future studies.

Base-editing screens with single-cell readouts across diverse hematopoietic lineages

As single-nucleotide variants can have effects that could be missed by specific functional readouts and might be limited to particular differentiation stages, we sought to enhance hematopoietic base-editing screens by obtaining single-cell transcriptional readouts.[22,26,49] For this purpose, we needed an approach that allowed detection of the sgRNA present in each cell as part of the cell's scRNA-seq readout. Among different approaches tested, we obtained the best results with a modified CROP-seq vector system, which embeds the sgRNA expression cassette sequence within a polyadenylated transcript that can be detected using scRNA-seq (Figures S3A and S3C–S3G; STAR Methods).[35] In contrast, we achieved minimal sgRNA detection in scRNA-seq reads with recently developed direct sgRNA capture approaches (Figure S3B).[50]

We next employed this scRNA-seq system to measure the perturbations caused by base editing—an approach we term Perturb(BE)-seq as it is adapted from other Perturb-seq approaches—but it employs base editing, rather than Cas9 nuclease or CRISPRi-mediated perturbation.[34–37] Before performing pooled screens, we first sought to demonstrate that we could (1) classify an sgRNA as having an effect despite incomplete editing efficiencies and (2) detect effects across all hematopoietic lineages. For these purposes, we applied the methods to HSPCs treated separately in four ways (Figure 3A): base editing with the sgRNA targeting the exon 1 splice donor site of CD33 (Figures 2B and 2C), base editing with two negative control sgRNAs (a NT sgRNA and a sgRNA targeting the *AAVS1* safe harbor locus), and Cas9-cutting with a positive control cutting sgRNA targeting CD33 (Figure 2C).[51]

We titrated down the base editor protein dosage to achieve ~10% editing efficiency (to mimic sgRNAs with lower efficiency, which would be encountered in pooled screens), edited HSPCs in the different ways mentioned above, and allowed them to differentiate into multiple lineages; we pooled the experiments prior to scRNA-seq (Figures 3A and 3B). In 96.77% of cells, the relevant sgRNA could be readily identified from scRNA-seq reads (Figures 3C, S3A, and S3C–S3G; STAR Methods). We also measured CD33 protein levels with a barcoded antibody in tandem with the scRNA-seq readouts (Figure S3I).

With respect to the sensitivity of the approach, we found that even at 10% editing efficiency we detected CD33 as the gene with the most significant reduction in expression (based on adjusted p value) in cells transduced with the CD33 splice donor 1 sgRNA vs. cells with the AAVS1 sgRNA (Figure 3D).

With respect to screening across hematopoietic lineages, we found robust reduction in CD33 protein levels across all lineages present (Figures 3B, 3E, and S3H). We detected significantly more cells with low CD33 protein levels across all lineages for both ABE and Cas9 CD33-editing sgRNAs (Figures 3E and S3J). These results suggest that the approach can effectively create and detect perturbations across the full spectrum of cell states in human hematopoiesis.

Base editor screens in primary human erythroid cells can decipher non-coding variant contributions to the regulation of HbF

We next explored the use of base editor screens in primary hematopoiesis to identify regulatory variants that affect gene expression, which has implications for both interpretation of naturally occurring variants and the design of therapeutic interventions. We focused on the regulation of fetal hemoglobin (HbF), for which increased expression can suppress the effects of sickle cell disease and β-thalassemia. BCL11A, a key repressor of HbF expression, was discovered based on human genetic studies, and multiple clinical trials are testing the therapeutic suppression of BCL11A.[52–55] In addition, the highly homologous promoters of *HBG1* and *HBG2* (which encode the γ-globin protein that forms HbF, when combined with α-globin), are also key targets for emerging genome-editing approaches to treat sickle cell disease and β-thalassemia.[56] Naturally occurring mutations in these non-coding regulatory regions of *HBG1/2* can cause elevations of HbF in adults, a condition termed hereditary persistence of HbF (HPFH), most often by creating or destroying binding sites for the transcription factors GATA1, KLF1, BCL11A, or ZBTB7A.

We focused on the *HBG1/2* promoter regions as they have been studied extensively in cell lines but not systematically in primary cells, which are the targets of clinical therapies.[53,57,58] We assessed the effects of 124 sgRNAs targeting a span of 300-base pairs upstream of the transcription start sites of *HBG1/2* in primary HSPCs and assayed cells as they underwent semi-synchronous erythroid differentiation[59] (Figure 4A; STAR Methods). To be able to detect the identity of perturbations encoded by lentiviral sgRNAs, we profiled erythroblasts prior to enucleation (Figure S4A; STAR Methods). Based on intracellular HbF levels as assayed by FACS, we analyzed two groups of erythroblasts: those expressing the top 30% of HbF (HbF^{high}) and those expressing the lowest 30% of HbF (HbF^{low}). Whereas FACS-based functional screens compared the distribution of sgRNAs in two extreme populations of cells (with high and low levels of HbF), Perturb(BE)-seq measures the expression of *HBG1/2* transcript levels in all profiled cells, which we reasoned might have greater sensitivity compared with the former approach. We profiled 76,961 single erythroblasts and identified sgRNAs with significantly different *HBG1/2* transcript distributions (STAR Methods; Figure S4B). Additionally, to evaluate and confirm the editing outcomes of the sgRNAs used in a screen, we applied single-cell genotyping of 8,388 erythroblasts with 30 primer pairs targeting the sgRNA sequence and the *HBG1/2* promoters, as well as other loci and controls

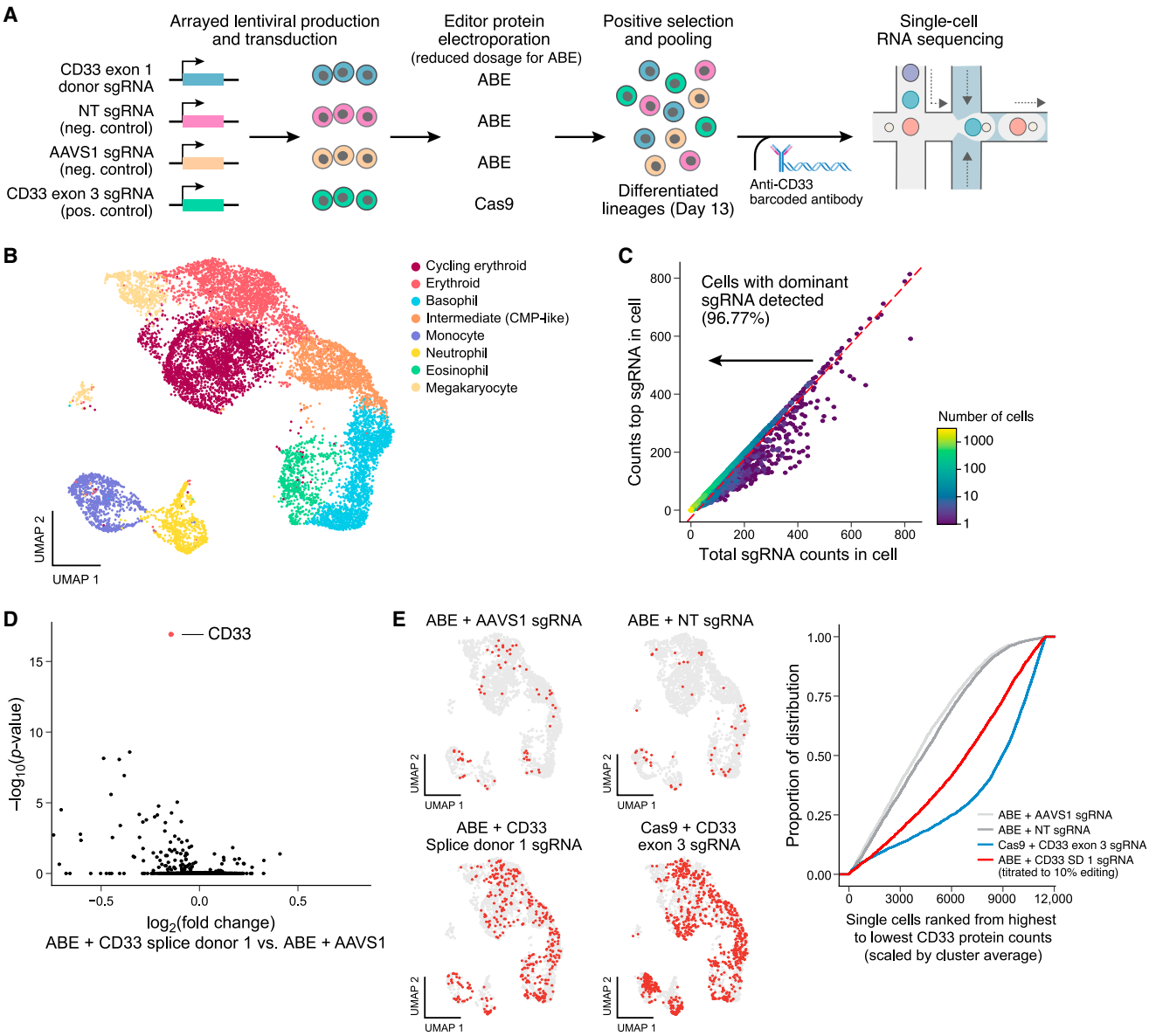


Figure 3. Capturing variant effects in the hematopoietic differentiation continuum with single-cell screens

(A) Schematic of the experimental design to detect single-cell perturbation effects using lentivirally transduced sgRNAs across primary human hematopoiesis. In this benchmarking experiment, lentiviruses were produced in an arrayed format given that the positive control condition was edited with a Cas9 nuclease and the others with adenine base editor (ABE).

(B) Uniform manifold approximation and projection (UMAP) of the different hematopoietic lineages arising after spontaneous HSPC differentiation *in vitro*. Hematopoietic lineages were assigned using the expression of known marker genes (Figure S3H).

(C) Scatterplot of the counts of the top sgRNA (measured using CROP-seq transcript counts detected following enrichment PCR) in each cell relative to the counts of all sgRNAs detected in that cell.

(D) Volcano plot of the transcriptome-wide $\log_2$(FC) and associated $-\log_{10}$(p values) between cells transduced with the CD33 splice donor 1 sgRNA or AAVS1 sgRNA.

(E) Left, UMAPs split by the identity of the transduced sgRNA. Red dots highlight the cells in the bottom 10% decile of CD33 protein counts (across all conditions), scaled by the average counts of each cluster from (B). Right, cumulative distribution of single-cell CD33 protein counts (scaled by cluster average) across the four experimental conditions.

See also Figure S3.

(Figures 4A and S4C; STAR Methods). Genotyping may be particularly valuable for screens in which it is important to connect sgRNAs to their mutational outcomes. While other approaches (such as PCR enrichment on scRNA-seq reads[61,62]) allow for variant calling in transcribed regions, genotyping is more general in that it can be applied to any genomic region,

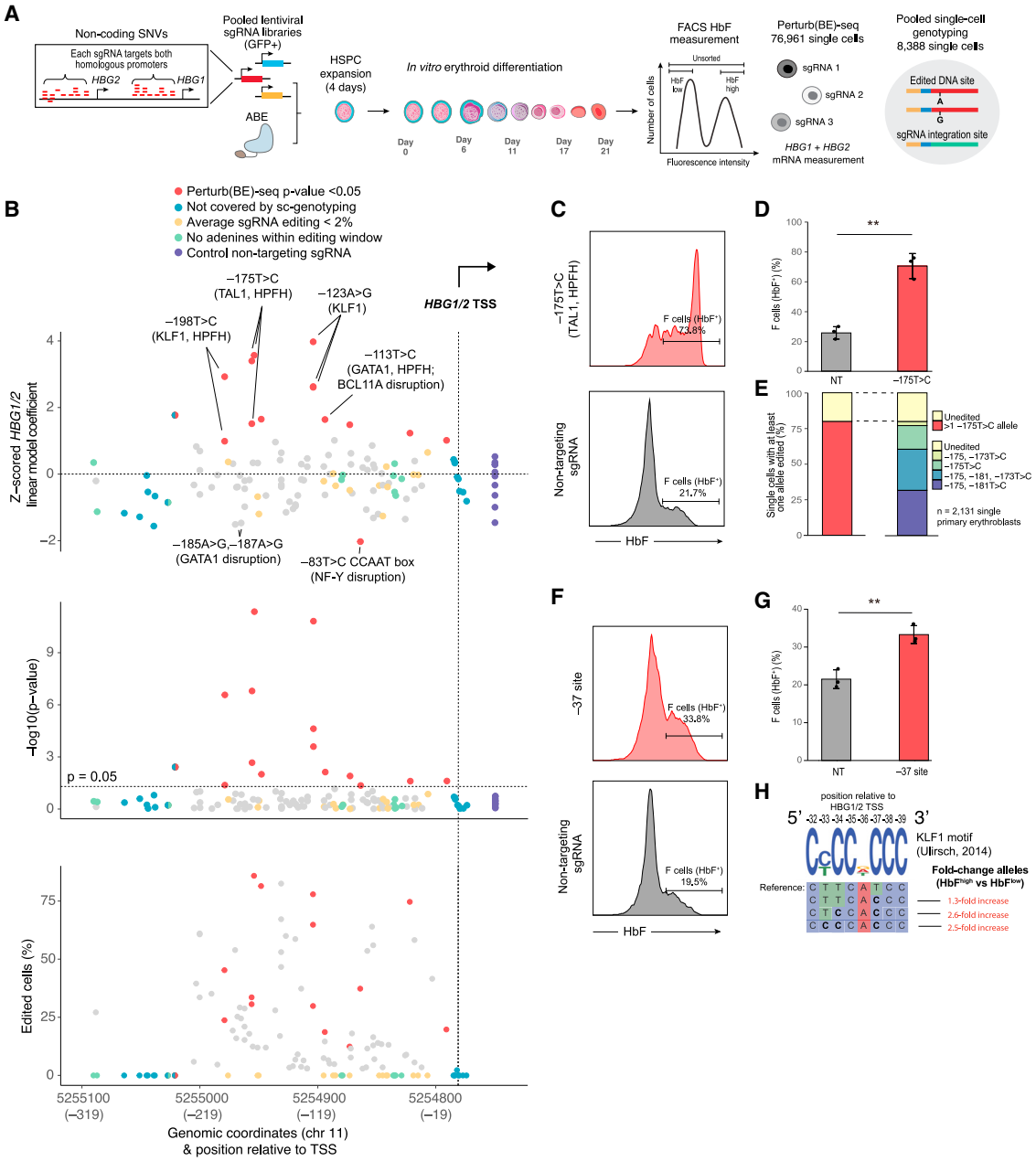


Figure 4. Dissection of the regulatory logic of erythroid-specific non-coding regions with base editor screens

(A) Schematic of pooled ABE8e screens targeting the *HBG1/2* promoters. Functional screens and Perturb(BE)-seq were performed on day 13 erythroblasts to capture sgRNA information prior to enucleation. Pooled single-cell genotyping was conducted on day 6 of erythroid differentiation.

(B) Top track, *Z* scored linear model coefficients for each sgRNA in the screen (STAR Methods). Plotted genomic coordinates display the most common edited nucleotide for each sgRNA. Middle track, p values from the linear model shown in the top track. Bottom track, percentage of edited single cells for each sgRNA in the screen.

(C) Representative flow-cytometry measurement of HbF levels in erythroid-differentiated HSPCs (day 14) treated with ABE precomplexed with an sgRNA targeting −175T>C or non-targeting sgRNA.

(D) Average percentage of HbF^+ cells (F cells) in erythroid-differentiated HSPCs (day 14) treated with ABE precomplexed with an sgRNA targeting the −175T>C or non-targeting sgRNA. 3 independent electroporations. Two-tailed unpaired t test.

(E) Pooled single-cell genotyping experiments of HSPCs treated with ABE precomplexed with an sgRNA targeting −175T>C. Left, percentage of single cells with at least one −175T>C edited allele. Right, percentage of single cells with at least one allele of −173T>C, −175T>C, or −181T>C, which reside within the editing window.

(F) Representative flow-cytometry measurement of HbF levels in erythroid-differentiated HSPCs (day 14) treated with ABE precomplexed with an sgRNA targeting the −37 site or non-targeting sgRNA.

(legend continued on next page)

including non-coding regions (as shown here for the *HBG1/2* promoters) and coding regions in lowly expressed genes or far from the universal primers attached to transcript ends in scRNA-seq protocols.

Both the functional screen and Perturb(BE)-seq identified three critical sites (with 2–3 independent sgRNAs per site) previously known to increase HbF among the top hits (Figures 4B, top and middle tracks and S4D): −123, −175, and −198 nucleotides upstream of the transcription start sites.[57,63–65] We also observed enrichment in HbF^{low} cells of two sgRNAs targeting the −185 GATA1 activator motif, a site that has been previously suggested to decrease HbF levels when mutated.[66] As anticipated, the Perturb(BE)-seq screen identified a larger number of hits, consistent with the increased power of this screen. Pooled single-cell genotyping enabled precise identification of the nucleotides edited by each sgRNA (Figures 4B, bottom track and S4E).

Given the clinical utility of HbF induction as a curative therapy for sickle cell disease and β-thalassemia, we sought to validate the sgRNAs targeting the −175T>C HPFH alteration, as this site was one of the strongest hits in the screen and results in one of the highest HbF levels reported in patients in the literature. This mutation is thought to create a TAL1 binding site and thereby increase transcriptional activity as a result.[63,64] Erythroid cells individually edited with one of the sgRNAs targeting this site resulted in 70.4% ± 8.4% HbF^{+} cells compared with 25.6% ± 4.2% in cells treated with a NT control sgRNA (Figures 4C and 4D). Quantitative real-time PCR confirmed an increased ratio of *HBG1/2* to *HBB* compared with control cells (Figure S4F). Single-cell genotyping of 2,131 edited primary erythroblasts showed editing of at least one −175T>C allele in 80% of cells, with a distribution of additional edits in the editing window (Figure 4E). Analysis of the top 5 predicted off-target sites revealed no detectable editing (Figure S4G). We confirmed normal differentiation and morphology of edited cells (Figures S4H and S4I). These results demonstrate the robustness of screens to identify key nucleotide alterations that could enable improved therapeutic strategies.

In addition to the sites found by the functional screen, the Perturb(BE)-seq screen identified a number of additional sites in the *HBG1/2* promoters. The −113A>G mutation lies in a region known to bind the HbF repressor BCL11A that is mutated in individuals with HPFH[66–68] (Figure 4B). We also observed reduced *HBG1/2* levels in cells with sgRNAs recreating −83T>C. This mutation lies within a CCAAT box previously shown to bind the NF-Y transcription factor, which acts as a transcriptional activator of *HBG1/2*.[66] We further investigated one of the hits targeting a previously unstudied site in the promoter, spanning a number of adenines centered around position −37. Individual validation of the sgRNA resulted in 30.7% ± 2.2% HbF^{+} cells compared with 19.8% ± 2.3% in control with 5.5% and 17.7% editing efficiencies at the −37 and −41 adenines, respectively (Figures 4F and 4G). Assessment of mRNA levels in erythroblasts confirmed the increased *HBG1/2* levels (Figure S4J). We confirmed normal differentiation and morphology of edited cells (Figures S4K and S4L). These mutations were predicted to create a *de novo* KLF1 binding motif that we surmised could underlie the observed HbF induction. Consistent with this, the genotype in FACS-purified HbF^{high} erythroblasts edited with a slightly offset and optimized sgRNA demonstrated an enrichment of mutations recreating a CACC box KLF1 motif (Figure 4H). This finding, along with the observations from the creation of a TAL1 binding site by the −175 mutation (Figures 4C–4E) exemplifies how the *HBG1/2* promoters are poised, but not optimized for maximal expression in adult erythroid cells and showcases strategies that could be employed to maximize expression of HbF for therapeutic purposes. More broadly, these approaches are likely to be valuable to systematically functionalize the many non-coding variants identified in human genetic studies in relevant cellular contexts.

Systematic mutagenesis of the master hematopoietic regulator GATA1 and its impacts on differentiation

We next sought to use Perturb(BE)-seq to enable structure-function mapping of how coding mutations across a gene affect cellular phenotypes across hematopoietic cell states and lineage transitions. We focused on GATA1 as an ideal test case, because it is a master hematopoietic zinc finger (ZF) transcription factor necessary for erythropoiesis, megakaryopoiesis, mast cell production, eosinophil differentiation, and basophil maturation.[69] Germline mutations in *GATA1* cause a number of different blood diseases, including Diamond-Blackfan anemia, congenital dyserythropoietic anemia, unlinked β-thalassemia, porphyria, myeloid malignancy predisposition, and thrombocytopenia, with some patients showing a combination of different phenotypes depending upon their mutation.[70–75] Additionally, somatic mutations in *GATA1* can drive the transient myeloproliferative disease and megakaryoblastic leukemia in Down syndrome or with acquired trisomy 21.[75,76] While distinct pathogenic germline and somatic alleles of *GATA1* continue to be identified, in most cases the precise phenotypes that will result from a specific mutation cannot be predicted. Moreover, many variants are of unclear significance to the underlying pathogenesis.[77]

We performed a systematic mutagenesis screen of *GATA1* in two genetically male (XY) donors (as *GATA1* is located in the X chromosome) with 514 sgRNAs targeting all exons and exon-intron boundaries (Figures 5A and S5A; STAR Methods). To capture the widest range of functional effects induced by mutations, cells were initially cultured in HSPC maintenance media for 4 days (STAR Methods) before transitioning the cells to semi-synchronous erythroid differentiation with scRNA-seq conducted on 278,675 single cells on days 2, 4, 7, and 9 of erythroid differentiation to capture the effects of mutations at different stages of erythroid maturation.[78,79]

(G) Average percentage of HbF^{+} cells (F cells) in erythroid-differentiated HSPCs (day 14) treated with ABE precomplexed with an sgRNA targeting the −37 site or non-targeting sgRNA. 3 independent electroporations. Two-tailed unpaired t test.
(H) Fold change of edited alleles between FACS-purified HbF^{high} and HbF^{low} erythroblasts edited by an sgRNA targeting the −37 site. The reference allele and the predicted *de novo* KLF1 motif that the editing generates are shown on top.[60]
See also Figure S4.

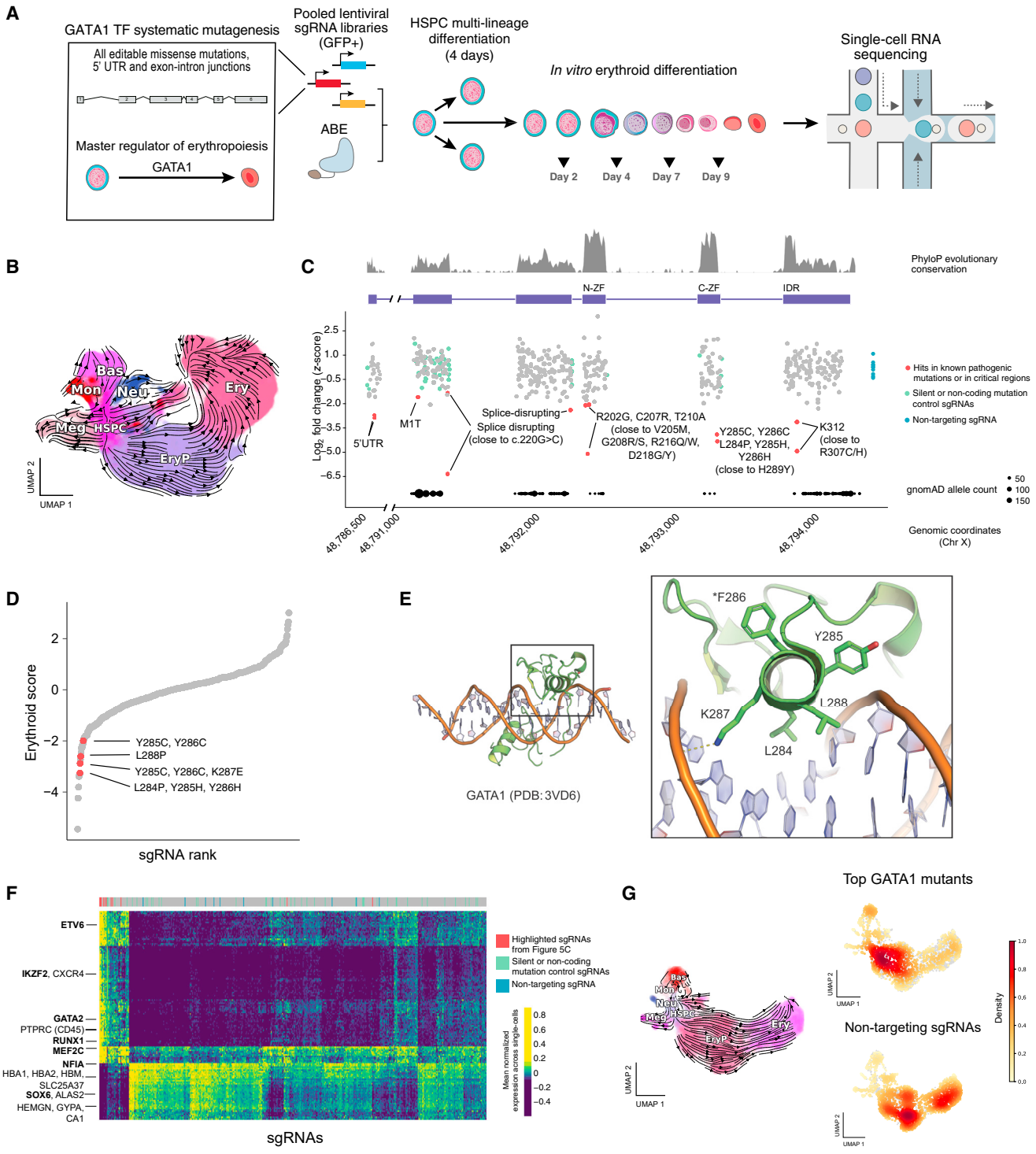


Figure 5. Systematic mutagenesis of the master hematopoietic transcription factor GATA1

(A) Schematic representation of the experiment targeting all editable missense mutations, exon-intron junctions, as well as the 5′ UTR and a subset of control mutations in *GATA1* (STAR Methods).

(B) UMAP of hematopoeitic cells with a dominant perturbation profiled at days 2, 4, 7, and 9 of erythroid differentiation in the *GATA1* screen. The streamline plot with the predicted RNA velocity flow projected in the UMAP space is overlaid. Hematopoietic lineages were assigned using known marker genes (Figure S5B).

(C) *Z* scored $\log_2$(FC) of sgRNA in cells sampled on day 9 of erythroid differentiation vs. transduced cells prior to electroporation, using bulk amplicon sequencing. Hits targeting previously known mutations or in critical regions of *GATA1* are highlighted, as well as control sgRNAs that were included to target silent mutations or

(legend continued on next page)

Based on gene expression, the resulting cells were a mixture of predominantly erythroid cells with some HSPCs, as well as neutrophilic, monocytic, megakaryocytic, and basophilic precursors (Figures 5B and S5B). We compared the representation of each sgRNA at the day 9 differentiation time point relative to the cells prior to electroporation to identify variants critical for GATA1 function (STAR Methods). We also ranked mutations by their specific impact on erythroid differentiation potential. We defined an "erythroid score" for each sgRNA as the *Z* scored proportion of cells observed in erythroid lineages vs. all lineages (STAR Methods), with low erythroid scores indicating a selective decrease in erythroid differentiation.

We identified hits across all exons, including known pathogenic mutations and uncharacterized mutations in critical regions such as the N- and C-terminal ZFs, as well as multiple splice sites (Figures 5C and S5C).[70,71,75,80–82] Intriguingly, we also observed two hits editing the same exon 1 5′ UTR nucleotide (GATA1 has one of the highest translational efficiencies among hematopoietic transcription factors and is particularly susceptible to reduced ribosomal levels[83]) and two hits in a highly conserved lysine in the intrinsically disordered region in the C terminus of GATA1, in which a recently described neighboring mutation causing congenital anemia results in altered transcriptional activity and chromatin occupancy.[77] In contrast, neither the sgRNAs recreating control non-pathogenic mutations nor the NT sgRNAs had *Z* scores < -1.1. Looking at sgRNAs with the lowest erythroid scores (which implies a selective depletion of these sgRNAs in erythroid cells), we observed a significant number targeting one of the critical α helices in the DNA-interacting C-terminal ZF domain (Figures 5D and S5D), which were not depleted in other lineages (Figure S5E). Specifically, point mutations L284P, Y285C/H, Y286C, K287E, and L288P likely destabilize the α helix and alter the interaction interface (Figure 5E).

Finally, we sought to gain insight into the downstream transcriptional effects arising from *GATA1* mutations. We identified differentially expressed genes shared among perturbed erythroid progenitors and precursors for the sgRNAs with the lowest erythroid scores compared with controls, and then clustered all sgRNAs by these genes (Figure 5F; STAR Methods).[84] We observed shared transcriptional responses among the most strongly depleted sgRNAs over the course of erythroid differentiation shown in Figure 5C: a decrease in terminal erythropoiesis genes (*HBA1*, *HBA2*, *HBM*, *SLC25A37*, *ALAS2*, *HEMGN*, *GYPA*, *CA1*, *SOX6*, and *NFIA*) and an increase in non-erythroid or early progenitor gene expression programs (*PTPRC*, *CXCR4*, *ETV6*, *IKZF2*, *RUNX1*, *GATA2*, and *MEF2C*). This strategy additionally identified sgRNAs that were not depleted over the course of erythroid differentiation but shared similar transcriptional responses, increasing the number of putative deleterious mutations identified in the screen, compared with the depletion analyses alone. Reconstruction of the predicted differentiation trajectories for cells transduced with these sgRNAs confirmed a block at the progenitor stages with impaired terminal differentiation (Figure 5G; STAR Methods). Taken together, our results nominate nucleotides critical for GATA1 function and provide mechanistic insights into their transcriptional and functional consequences over the course of hematopoietic differentiation.

Hematopoietic base editor screens help classify the pathogenicity of VUS

With more widespread application of clinical sequencing, large numbers of variants of unknown significance (VUS) are increasingly being discovered in patients, but determining whether they are causal remains challenging and limits effective clinical decision-making.[85] We reasoned that the data from systematic Perturb(BE)-seq screens for GATA1 could enable functional assessment of VUSs. We identified a male patient with congenital hypoplastic anemia, hemizygous for a reported VUS (c.220+2T>C) in the second exon-intron junction of *GATA1* (Figure 6A; STAR Methods). Bone marrow aspirates from this patient revealed notable erythroid hypoplasia and dyserythropoiesis (Figure S6A). We have previously reported a nearby c.220G>C synonymous mutation in two siblings with Diamond-Blackfan anemia, which affected splicing and resulted in the exclusive production of the short isoform of GATA1, called GATA1s, which lacks the transactivation domain (TAD).[70] However, the pathogenicity of the c.220+2T>C mutation is unclear.

Because this mutation was predicted to be recreated by two sgRNAs in our screen, we checked the pooled single-cell genotyping data and found that these sgRNAs actually produced the mutation in 62.9% and 48.3% of cells carrying the respective sgRNA (Figure 6B). In the Perturb(BE)-seq screen, cells transduced with the sgRNAs targeting c.220+2T>C depleted over the course of erythroid differentiation, were among the strongest transcriptional perturbations and were enriched

non-coding regions, and non-targeting controls. PhyloP evolutionary conservation scores and Genome Aggregation Database (gnomAD) allele counts at each position are included for reference (STAR Methods).

(D) *Z* scored ratio between cells in erythroid lineages and non-erythroid lineages for each sgRNA (erythroid score). Screen hits on the GATA1 C-terminal zinc finger are highlighted (Figure 5D).

(E) Crystal structure of murine GATA1 zinc fingers interacting with DNA (from https://doi.org/10.2210/pdb3VD6/pdb). Erythroid score screen hits presented in (D) cluster in the DNA-interacting α helix from the C-terminal zinc finger, and are labeled in the figure. The sequence is identical to human GATA1 zinc fingers with the exception of F286, which is Y286 in human (*F286). Edited amino acids were annotated using pooled single-cell genotyping.

(F) Heatmap of the mean expression levels of differentially expressed genes between sgRNAs with the lowest functional scores and non-targeting sgRNAs. Hierarchical clustering was performed both on the displayed genes and sgRNAs. Hits highlighted in (C) clustered together, as well as additional sgRNAs that shared a similar transcriptional response. A selection of relevant differentially expressed genes is highlighted, with transcription factors in bold.

(G) Left, streamline plot with the predicted RNA velocity flow projected in UMAP space using cells with the top *GATA1* perturbations and NT controls. Transitions of HSPCs into the different lineages are observed, with most of the cells giving rise to erythroid progenitors (EryPs) and erythroid precursors (Ery). Right, density plots in cells with the top *GATA1* perturbations highlight a block at the progenitor stages with impaired terminal differentiation compared with non-targeting controls.

See also Figure S5.

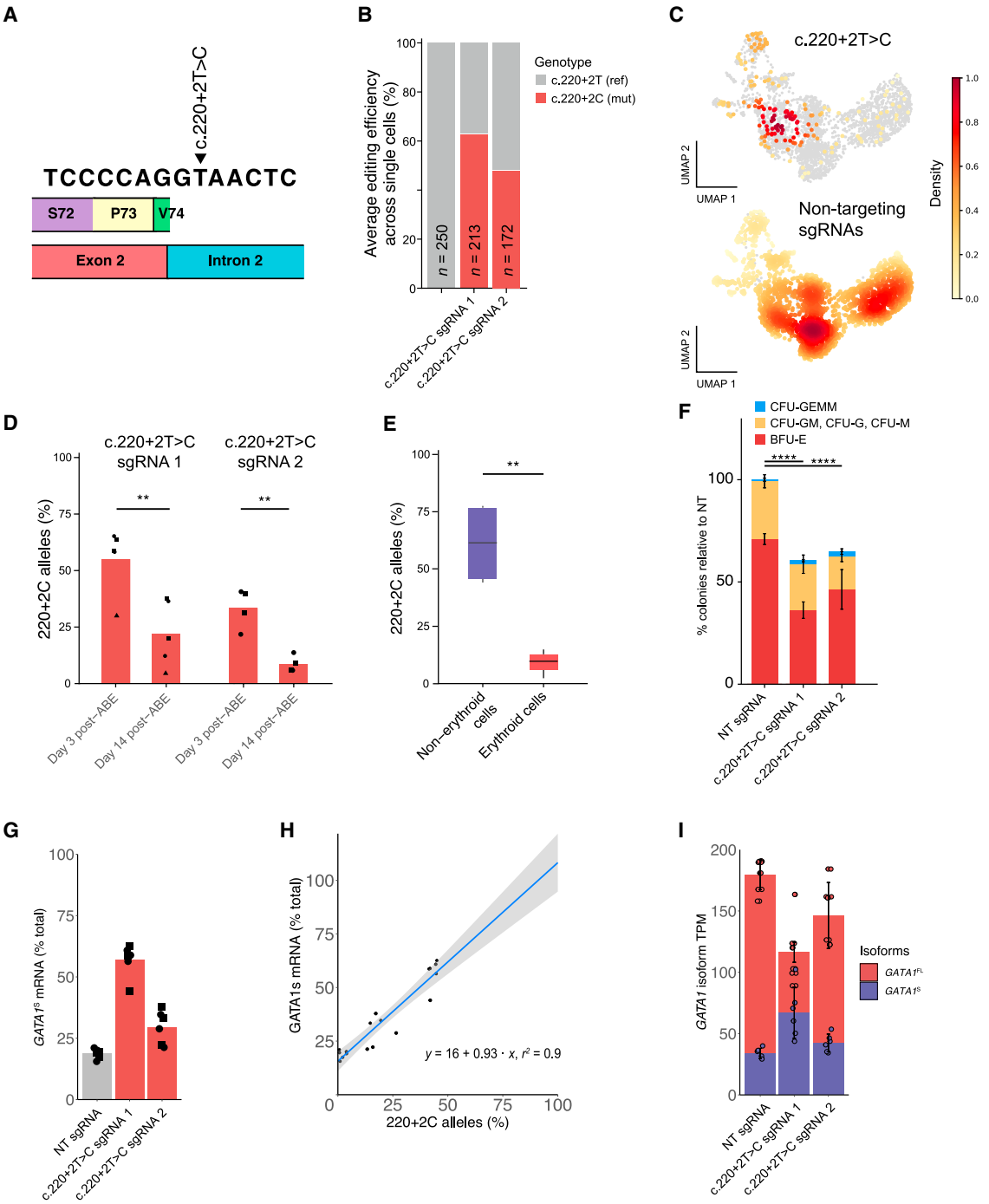


Figure 6. Defining pathogenicity of GATA1 variants in patients through base-editing screens

(A) Schematic of the GATA1 VUS (c.220+2T>C) in the second exon-intron junction in a patient with congenital anemia.

(B) Pooled single-cell genotyping of cells infected with the library in a healthy XY donor confirms editing by both sgRNAs targeting the patient VUS. Shown are the average editing efficiencies across all single cells with sgRNAs targeting c.220+2T>C. The number of single cells genotyped for each sgRNA is overlaid on the bar plots.

(C) UMAP density plots highlighting cells with the c.220+2T>C and 10 non-targeting sgRNAs. This mutation causes a block at the progenitor stages with impaired terminal differentiation, compared with non-targeting controls.

(D) Percentage of alleles edited with the c.220+2T>C mutation 3 and 14 days following electroporation with ABE and chemically modified sgRNAs in HSPCs subject to erythroid differentiation. Each dot represents independent electroporations, and the shape of the dot represents different HSPC donors. Two-tailed paired t test.

(legend continued on next page)

among non-erythroid lineages compared with the erythroid cells, suggesting a block in erythroid differentiation resulting from this mutation (Figures 5F, 6C, and S6B). We validated these findings using precomplexed base editors and sgRNAs to introduce the mutation in HSPCs from healthy donors (STAR Methods). Cells carrying the mutation were depleted over the course of erythroid differentiation and showed impaired erythroid differentiation potential (Figures 6D and S6C); in contrast, we found enrichment of the mutation in edited cells in the CD71−CD235a− fraction of non-erythroid cells (Figure 6E). We also performed methylcellulose colony-forming assays on base-edited HSPCs, which showed reduced erythroid colony formation, with preserved myeloid differentiation (Figure 6F). Full-length transcriptomic analysis of differentiating erythroid precursors on day 9 post-electroporation revealed an increase in the percentage of the *GATA1* short isoform in cells edited with the c.220+2T>C sgRNAs (which was proportional to the percent 220+2C edited alleles, Figures 6G and 6H), concomitant with a decrease in the absolute levels of the *GATA1* full-length isoform (Figure 6I). This demonstrates that the 220+2T>C mutation results in preferential splicing of *GATA1* to the short, rather than full-length isoform, thereby perturbing erythropoiesis and causing hypoplastic anemia (Figure S6D). Collectively, our findings reveal how a Perturb(BE)-seq screen in primary hematopoietic cells can be effectively used to rapidly identify a pathogenic variant that had previously been of uncertain significance.

Expanding the screening tool kit with CBEs

Our studies above employed ABEs, which can recreate many but not all variants. For example, while examining additional *GATA1* VUSs, we identified a second individual with congenital hypoplastic anemia and a distinct VUS (c.218C>T, resulting in P73L) four nucleotides away from the pathogenic c.220+2T>C variant assessed in our ABE Perturb(BE)-seq screen (STAR Methods; Figures 7A and S7A). This specific variant could not be introduced in our initial screen, but is targetable using CBEs.

To examine the utility of other base editors in human hematopoiesis, we purified an evolved cytosine deaminase, evoFERNY,[86] and added a flexible NG PAM for broader targeting (Figure S7B; STAR Methods). We obtained fractions of modified alleles of around 50% in primary HSPCs using precomplexed chemically synthesized sgRNAs with the evoFERNY protein (Figure S7C). By using a similar strategy as for the c.220+2T>C variant, we recreated the unclassified c.218C>T variant in human HSPCs with two different sgRNAs and induced erythroid differentiation. We observed complete depletion of alleles resulting in c.218C>T (P73L) during erythroid differentiation (Figure 7B). Colony-forming assays confirmed that the mutation selectively impaired erythroid differentiation without compromising other myeloid lineages (Figure 7C). This result demonstrates our ability to utilize additional editors to model pathogenic mutations at endogenous loci, using similar approaches as for ABE8e in human HSPCs; together, these editors expand the spectrum of variants that can be recreated, while additional changes are made possible with emerging editors.[12–16]

Given our success in recreating a single pathogenic variant identified in a patient, we sought to assess the ability to conduct functional screens. Therefore, we performed another systematic mutagenesis screen of *CD33* splice sites, as was done with ABE8e, but with evoFERNY protein instead. The results of this screen were robust and the top hit of the screen was the exact same sgRNA targeting the exon 1 donor splice site, but with a different (C>T) mutation—underscoring the relevance of this splice site for CD33 expression (Figure 7D). Notably, the screen also identified the exon donor 2 site noted in our ABE8e screen, and additionally nominated the splice acceptors in exons 3 and 4. The exon 3 splice-acceptor site was not seen in the earlier screen, likely because ABEs tend to be less effective than CBEs at editing splice-acceptor motifs,[43] while the exon 4 splice-acceptor site could not be targeted in the ABE screen. Taken together, these results reinforce the modularity of our approaches to readily expand the type and number of variants targeted by replacing the editor with different purified genome-editing proteins.

DISCUSSION

The inability to study the effects of large numbers of single-base substitutions in primary human cells has been a major bottleneck in understanding cellular function and disease pathogenesis. Here, we introduce a platform for massively parallel base editing and apply it to primary human HSPCs and their differentiated progeny.

We focus on describing the robustness of the approach with proof-of-principle experiments that provide important biological insights. These experiments include a gene-editing strategy for improving leukemia immunotherapy, the characterization of large numbers of variants modulating HbF levels in primary hematopoietic cells, and a systematic mutagenesis screen of *GATA1* variants that enables the classification of disease-causal

(E) Percentage of alleles edited with the c.220+2T>C mutation 9 days following electroporation with ABE and chemically modified sgRNAs in sorted erythroid cells and non-erythroid cells. Boxplots summarize data from two HSPC donors and the two c.220+2T>C sgRNAs. Two-tailed paired t test.
(F) Methylcellulose colony-forming assays from XY healthy donors edited with the c.220+2T>C mutation using ABE protein and chemically modified sgRNAs. The percentage of each colony type is normalized to the total number of colonies electroporated with non-targeting sgRNAs. The error bars represent the standard deviation in the normalized number of colonies across two donors with three technical replicates each for each sgRNA. Two-tailed unpaired t test.
(G) Percentage of *GATA1* short isoform (*GATA1s*) mRNA with respect to the total *GATA1* mRNA transcripts in differentiating erythroid precursors on day 9 post-electroporation edited with the c.220+2T>C sgRNAs or NT control. Each dot represents an independent electroporation, and the shape of the dot represents different HSPC donors.
(H) Percentage of *GATA1* short isoform (*GATA1s*) mRNA with respect to the total *GATA1* mRNA transcripts in differentiating erythroid precursors on day 9 post-electroporation edited with the c.220+2T>C sgRNAs or NT control, as a function of the editing efficiency.
(I) Transcripts per million of *GATA1* short (*GATA1s*) and *GATA1* full-length isoform (*GATA1FL*) isoforms in differentiating erythroid precursors on day 9 post-electroporation.
See also Figure S6.

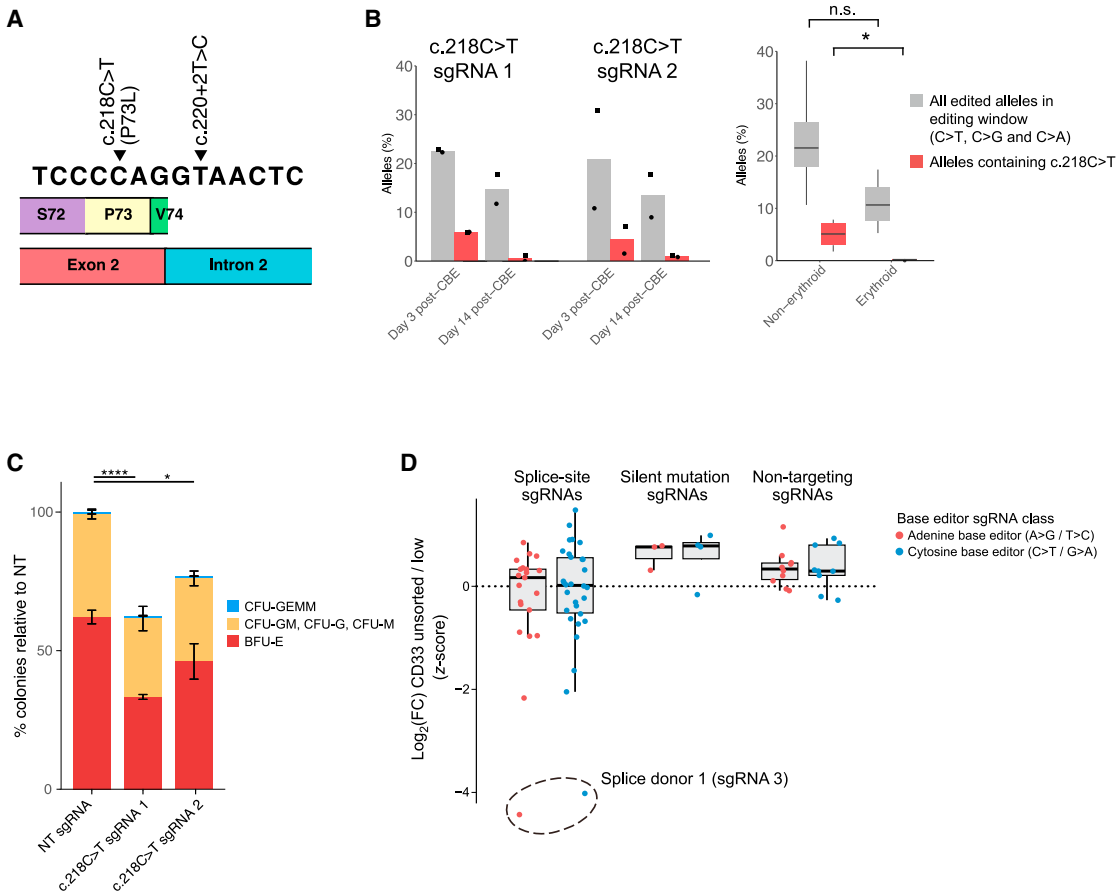


Figure 7. Expanding the screening tool kit with cytosine base editors

(A) Schematic of the two GATA1 VUSs validated in this paper.

(B) Left, percentage of alleles bearing the c.218C>T mutation (red bars) or other C>A and C>G mutations, as well as a small fraction of indels (gray bars) in edited HSPCs subjected to erythroid differentiation from two XY donors. Data are shown for two contiguous, different sgRNAs targeting c.218C>T. Each dot represents independent electroporations, and the shape of the dot represents different HSPC donors.

(C) Methylcellulose colony-forming assays in two healthy donors edited with sgRNAs targeting the c.218C>T mutation. The error bars represent the standard deviation in the number of colonies (normalized to NT) across two donors with three technical replicates for each donor and sgRNA.

(D) *Z* scored $\log_2$(FC) in sgRNA reads between HSPCs with the bottom 10% CD33 levels and the unsorted population, for both orthogonal adenine and cytosine base editor screens. Each dot represents an sgRNA.

See also Figure S7.

variants in this gene. Notably, this approach makes it possible to study the effects of variants in nearly all hematopoietic lineages, given the ability to readily differentiate human HSPCs *in vitro* to most blood and immune cell lineages—including erythroid cells, megakaryocytes, basophils, eosinophils, neutrophils, monocytes, dendritic cells, innate lymphoid cells, NK cells, T lymphocytes, and B lymphocytes[59,87–95]— and to recognize the cell types by functional markers and/or gene expression. Efficient editing and sensitive readouts allow screening for complex phenotypes.

The ability to make base substitutions in endogenous loci has many advantages, compared with strategies based on overexpressing mutant cDNAs from plasmids.[96] It expands the types of genomic elements that can be interrogated at base-pair resolution to include splice sites and non-coding regions. The approach will be particularly valuable for experimental evaluation of large numbers of single-nucleotide variants being found through human genome-wide association studies and VUSs being identified through clinical sequencing of patients to identify pathogenic mutations.[97] Importantly, the approach allows the careful study of mutated protein function at physiological levels of expression and in primary cells. The strategy also makes it possible to revert variants in patient-derived cells to study their functional consequences, such as testing whether a blood disorder phenotype can be reversed through editing of a particular candidate mutation in HSCs.

Unlike pooled CRISPR-Cas9 screens, in which a gene knockout can be reliably evaluated by using many different sgRNAs against a gene to overcome variability in cutting efficiency, only a few sgRNAs are capable of engineering a specific base substitution. The inability to filter out sgRNAs with poor editing efficiencies might result in false negatives. To minimize the

risk of false negatives, we employ PAM-flexible editors to increase the number of guides per variant, and we employ pooled single-cell genotyping to directly assess editing outcomes to augment these screens.

Our approach is modular and allows for ready replacement with other genome-editing tools, such as by Cas9 enzymes with expanded PAM compatibility.[98,99] Combining the expanding spectrum of genome editors (including alternate base editors, prime editors, or epigenetic modifiers[14,16,99–104]) and the ability to edit primary cells significantly advance the opportunities for targeted genome manipulation of human hematopoietic and other cells. Moreover, an advantage of editor-protein electroporation in tandem with lentiviral sgRNA library delivery is the lower likelihood of observing off-target effects given the shorter half-life of the electroporated protein, compared with mRNA or plasmid-based delivery approaches.[51,65,105–107] This improves the ability to translate screening results to therapeutic applications, by identifying the most efficient sgRNAs with the same editing strategies and in the same cell types that will be the target of curative therapies for human diseases, such as HSCs that can enable long-term reconstitution of hematopoiesis.

Potential improvements to the method may include the use of multiple sgRNA cassettes to enable multiplexed combinatorial screens, screens in patient-derived cells, and *in vivo* transplantation of human HSPCs transduced with pooled libraries to study variant dynamics with long-term hematopoiesis. We envision that the strategies introduced in this paper will enable a fine-grained understanding of how genetic variation predisposes to human disease and will provide important insights into the molecular logic of human hematopoiesis and other primary cell systems.

Limitations of the study

While the screening approaches we have described enable rich information of cell state at single-cell resolution via Perturb(BE)-seq and pooled genotyping of the endogenous loci targeted in the screen, these readouts are not coupled. Given the critical information obtained from transcriptomic readouts and limitations of existing approaches,[108] future advances should seek to incorporate methods that simultaneously enable DNA genotyping and gene expression readouts at high throughput for the same single cells in large pooled screens. This will boost statistical power and enable a more direct dissection of the phenotypic consequences of the individual alleles introduced by the same sgRNA. In addition, while we have demonstrated the value of base editing with ABEs and CBEs, additional base editors that have more recently been developed should be tested in this platform to recreate other types of variants.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2023.03.035.

ACKNOWLEDGMENTS

We thank members of the Sankaran lab, as well as Thouis Jones, Judhajeet Ray, Arnav Mehta, Jesus M. Chaves, Marina R. de Haro, Nadia Eckert, Sunil Nagpal, and Erik Ingelsson for assistance with protocols and discussions regarding this work. We are also grateful to Leslie Gaffney for help with graphics, to Mahsa Khanlari for assistance with pathology sample imaging, and to Jenna Sternberg and Susan Richards for research support and valuable comments. J.D.M.-R. is supported by fellowships from La Caixa Foundation (ID 100010434), the Rafael del Pino Foundation, and the American Society of Hematology. This work was supported by a GSK-Broad Institute research collaboration (V.G.S.), the New York Stem Cell Foundation (V.G.S.), the Edward P. Evans Foundation (V.G.S.), a gift from the Lodish Family to Boston Children's Hospital (V.G.S.), and National Institutes of Health (NIH) grants

R01 DK103794, R01 CA265726, R01 HL146500 (V.G.S.). G.A.N. receives support from NIH grant K99 HL163805. D.R.L. and G.A.N. acknowledge support from NIH awards U01 AI142756, RM1 HG009490, R35 GM118062, the Bill and Melinda Gates Foundation, and the Howard Hughes Medical Institute. X.Q. acknowledges funding from CZI and the Impetus Longevity Grant. V.G.S. is a New York Stem Cell-Robertson Investigator.

AUTHOR CONTRIBUTIONS

Conceptualization, J.D.M.-R. and V.G.S.; methodology, J.D.M.-R. and V.G.S.; formal analysis, J.D.M.-R., M.P., X.Q., A.M.R.-E., C.W., E.S.L., and V.G.S.; software, J.D.M.-R. and M.P.; investigation, J.D.M.-R., N.C., E.I.G., A.C., L.W., S.J., T.L., S.B., and G.M.; resources, S.C., G.A.N., A.A., D.R.L., M.W.W., K.S., J.H.O., R.M.F., R.J.X., D.E.K., and E.S.L.; writing – original draft, J.D.M.-R. and V.G.S.; writing – review & editing, J.D.M.-R., E.S.L., and V.G.S. with input from all authors; supervision, V.G.S.

DECLARATION OF INTERESTS

D.R.L. and G.A.N. have filed patent applications on gene-editing technologies through the Broad Institute of MIT and Harvard. D.R.L. is a consultant and equity owner of Beam Therapeutics, Pairwise Plants, Prime Medicine, Chroma Medicine and Nvelop Therapeutics, companies that use or deliver genome editing or genome engineering technologies. R.J.X. is the co-founder of Jnana Therapeutics and Celsius Therapeutics, the director of Moonlake Immunotherapeutics and a scientific advisory board member to Nestle, all unrelated to the present work. V.G.S. serves as an advisor to and/or has equity in Branch Biosciences, Ensoma, Novartis, Forma, and Cellarity, all unrelated to the present work.

INCLUSION AND DIVERSITY

One or more of the authors of this paper self-identifies as an underrepresented ethnic minority in their field of research or within their geographical location. One or more of the authors of this paper self-identifies as a gender minority in their field of research. One or more of the authors of this paper self-identifies as a member of the LGBTQIA+ community. We support inclusive, diverse, and equitable conduct of research.

Received: October 13, 2022
Revised: February 26, 2023
Accepted: March 30, 2023
Published: May 2, 2023

REFERENCES

1. Nandakumar, S.K., Liao, X., and Sankaran, V.G. (2020). In the blood: connecting variant to function in human hematopoiesis. Trends Genet. *36*, 563–576. https://doi.org/10.1016/j.tig.2020.05.006.
2. Uddin, F., Rudin, C.M., and Sen, T. (2020). CRISPR gene therapy: applications, limitations, and implications for the future. Front. Oncol. *10*, 1387. https://doi.org/10.3389/fonc.2020.01387.
3. Leibowitz, M.L., Papathanasiou, S., Doerfler, P.A., Blaine, L.J., Sun, L., Yao, Y., Zhang, C.Z., Weiss, M.J., and Pellman, D. (2021). Chromothripsis as an on-target consequence of CRISPR-Cas9 genome editing. Nat. Genet. *53*, 895–905. https://doi.org/10.1038/s41588-021-00838-7.
4. Cullot, G., Boutin, J., Toutain, J., Prat, F., Pennamen, P., Rooryck, C., Teichmann, M., Rousseau, E., Lamrissi-Garcia, I., Guyonnet-Duperat, V., et al. (2019). CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations. Nat. Commun. *10*, 1136. https://doi.org/10.1038/s41467-019-09006-2.
5. Schiroli, G., Conti, A., Ferrari, S., Della Volpe, L., Jacob, A., Albano, L., Beretta, S., Calabria, A., Vavassori, V., Gasparini, P., et al. (2019). Precise Gene Editing Preserves Hematopoietic Stem Cell Function following Transient p53-Mediated DNA Damage Response. Cell Stem Cell *24*, 551–565.e8. https://doi.org/10.1016/j.stem.2019.02.019.
6. Tao, J., Wang, Q., Mendez-Dorantes, C., Burns, K.H., and Chiarle, R. (2022). Frequency and mechanisms of LINE-1 retrotransposon insertions at CRISPR/Cas9 sites. Nat. Commun. *13*, 3685. https://doi.org/10.1038/s41467-022-31322-3.
7. Richter, M.F., Zhao, K.T., Eton, E., Lapinaite, A., Newby, G.A., Thuronyi, B.W., Wilson, C., Koblan, L.W., Zeng, J., Bauer, D.E., et al. (2020). Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. Nat. Biotechnol. *38*, 883–891. https://doi.org/10.1038/s41587-020-0453-z.
8. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. *42*, D980–D985. https://doi.org/10.1093/nar/gkt1113.
9. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44*, D862–D868. https://doi.org/10.1093/nar/gkv1222.
10. Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I., and Liu, D.R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature *551*, 464–471. https://doi.org/10.1038/nature24644.
11. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420–424. https://doi.org/10.1038/nature17946.
12. Tong, H., Wang, X., Liu, Y., Liu, N., Li, Y., Luo, J., Ma, Q., Wu, D., Li, J., Xu, C., et al. (2023). Programmable A-to-Y base editing by fusing an adenine base editor with an N-methylpurine DNA glycosylase. Nat. Biotechnol. https://doi.org/10.1038/s41587-022-01595-6.
13. Rees, H.A., and Liu, D.R. (2018). Base editing: precision chemistry on the genome and transcriptome of living cells. Nat. Rev. Genet. *19*, 770–788. https://doi.org/10.1038/s41576-018-0059-1.
14. Zhao, D., Li, J., Li, S., Xin, X., Hu, M., Price, M.A., Rosser, S.J., Bi, C., and Zhang, X. (2021). Glycosylase base editors enable C-to-A and C-to-G base changes. Nat. Biotechnol. *39*, 35–40. https://doi.org/10.1038/s41587-020-0592-2.
15. Chen, S., Liu, Z., Lai, L., and Li, Z. (2022). Efficient C-to-G base editing with improved target compatibility using engineered deaminase-nCas9 fusions. CRISPR J. *5*, 389–396. https://doi.org/10.1089/crispr.2021.0124.
16. Kurt, I.C., Zhou, R., Iyer, S., Garcia, S.P., Miller, B.R., Langner, L.M., Grünewald, J., and Joung, J.K. (2021). CRISPR C-to-G base editors for inducing targeted DNA transversions in human cells. Nat. Biotechnol. *39*, 41–46. https://doi.org/10.1038/s41587-020-0609-x.
17. Cuella-Martin, R., Hayward, S.B., Fan, X., Chen, X., Huang, J.W., Taglialatela, A., Leuzzi, G., Zhao, J., Rabadan, R., Lu, C., et al. (2021). Functional interrogation of DNA damage response variants with base editing screens. Cell *184*, 1081–1097.e19. https://doi.org/10.1016/j.cell.2021.01.041.
18. Hanna, R.E., Hegde, M., Fagre, C.R., DeWeirdt, P.C., Sangree, A.K., Szegletes, Z., Griffith, A., Feeley, M.N., Sanson, K.R., Baidi, Y., et al. (2021). Massively parallel assessment of human variants with base editor screens. Cell *184*, 1064–1080.e20. https://doi.org/10.1016/j.cell.2021.01.012.
19. Cheng, L., Li, Y., Qi, Q., Xu, P., Feng, R., Palmer, L., Chen, J., Wu, R., Yee, T., Zhang, J., et al. (2021). Single-nucleotide-level mapping of DNA regulatory elements that control fetal hemoglobin expression. Nat. Genet. *53*, 869–880. https://doi.org/10.1038/s41588-021-00861-8.
20. Xu, P., Liu, Z., Liu, Y., Ma, H., Xu, Y., Bao, Y., Zhu, S., Cao, Z., Wu, Z., Zhou, Z., et al. (2021). Genome-wide interrogation of gene functions through base editor screens empowered by barcoded sgRNAs. Nat. Biotechnol. *39*, 1403–1413. https://doi.org/10.1038/s41587-021-00944-1.

21. Sánchez-Rivera, F.J., Diaz, B.J., Kastenhuber, E.R., Schmidt, H., Katti, A., Kennedy, M., Tem, V., Ho, Y.-J., Leibold, J., Paffenholz, S.V., et al. (2022). Base editing sensor libraries for high-throughput engineering and functional analysis of cancer-associated single nucleotide variants. Nat. Biotechnol. *40*, 862–873. https://doi.org/10.1038/s41587-021-01172-3.

22. Yu, F., Cato, L.D., Weng, C., Liggett, L.A., Jeon, S., Xu, K., Chiang, C.W.K., Wiemels, J.L., Weissman, J.S., de Smith, A.J., et al. (2022). Variant to function mapping at single-cell resolution through network propagation. Nat. Biotechnol. *40*, 1644–1653. https://doi.org/10.1038/s41587-022-01341-y.

23. Freimer, J.W., Shaked, O., Naqvi, S., Sinnott-Armstrong, N., Kathiria, A., Garrido, C.M., Chen, A.F., Cortez, J.T., Greenleaf, W.J., Pritchard, J.K., et al. (2022). Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of immune networks. Nat. Genet. *54*, 1133–1144. https://doi.org/10.1038/s41588-022-01106-y.

24. Yu, K.-R., Corat, M.A.F., Metais, J.-Y., and Dunbar, C.E. (2016). 564. The cytotoxic effect of RNA-guided endonuclease Cas9 on human hematopoietic stem and progenitor cells (HSPCs). Mol. Ther. *24*, S225–S226. https://doi.org/10.1016/S1525-0016(16)33372-X.

25. Lattanzi, A., Meneghini, V., Pavani, G., Amor, F., Ramadier, S., Felix, T., Antoniani, C., Masson, C., Alibeu, O., Lee, C., et al. (2019). Optimization of CRISPR/Cas9 delivery to human hematopoietic stem and progenitor cells for therapeutic genomic rearrangements. Mol. Ther. *27*, 137–150. https://doi.org/10.1016/j.ymthe.2018.10.008.

26. Liggett, L.A., and Sankaran, V.G. (2020). Unraveling hematopoiesis through the lens of genomics. Cell *182*, 1384–1400. https://doi.org/10.1016/j.cell.2020.08.030.

27. Aiuti, A., Pasinelli, F., and Naldini, L. (2022). Ensuring a future for gene therapy for rare diseases. Nat. Med. *28*, 1985–1988. https://doi.org/10.1038/s41591-022-01934-9.

28. Wagenblast, E., Araújo, J., Gan, O.I., Cutting, S.K., Murison, A., Krivdova, G., Azkanaz, M., McLeod, J.L., Smith, S.A., Gratton, B.A., et al. (2021). Mapping the cellular origin and early evolution of leukemia in Down syndrome. Science *373*, eabf6202. https://doi.org/10.1126/science.abf6202.

29. Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.H., Raffield, L.M., Tardaguila, M., Huffman, J.E., et al. (2020). The polygenic and monogenic basis of blood traits and diseases. Cell *182*, 1214–1231.e11. https://doi.org/10.1016/j.cell.2020.08.008.

30. Claussnitzer, M., Cho, J.H., Collins, R., Cox, N.J., Dermitzakis, E.T., Hurles, M.E., Kathiresan, S., Kenny, E.E., Lindgren, C.M., MacArthur, D.G., et al. (2020). A brief history of human disease genetics. Nature *577*, 179–189. https://doi.org/10.1038/s41586-019-1879-7.

31. Dokal, I., Tummala, H., and Vulliamy, T. (2022). Inherited bone marrow failure in the pediatric patient. Blood *140*, 556–570. https://doi.org/10.1182/blood.2020006481.

32. Tangye, S.G., Al-Herz, W., Bousfiha, A., Cunningham-Rundles, C., Franco, J.L., Holland, S.M., Klein, C., Morio, T., Oksenhendler, E., Picard, C., et al. (2022). Human inborn errors of immunity: 2022 Update on the Classification from the International Union of Immunological Societies Expert Committee. J. Clin. Immunol. *42*, 1473–1507. https://doi.org/10.1007/s10875-022-01289-3.

33. Chen, M.H., Raffield, L.M., Mousas, A., Sakaue, S., Huffman, J.E., Moscati, A., Trivedi, B., Jiang, T., Akbari, P., Vuckovic, D., et al. (2020). Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell *182*, 1198–1213.e14. https://doi.org/10.1016/j.cell.2020.06.045.

34. Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single cell RNA profiling of pooled genetic screens. Cell *167*, 1853–1866.e17. https://doi.org/10.1016/J.CELL.2016.11.038.

35. Datlinger, P., Rendeiro, A.F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L.C., Kuchler, A., Alpar, D., and Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. Nat. Methods *14*, 297–301. https://doi.org/10.1038/nmeth.4177.

36. Adamson, B., Norman, T.M., Jost, M., Cho, M.Y., Nuñez, J.K., Chen, Y., Villalta, J.E., Gilbert, L.A., Horlbeck, M.A., Hein, M.Y., et al. (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. Cell *167*, 1867–1882.e21. https://doi.org/10.1016/j.cell.2016.11.048.

37. Jaitin, D.A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T.M., Tanay, A., van Oudenaarden, A., and Amit, I. (2016). Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. Cell *167*, 1883–1896.e15. https://doi.org/10.1016/j.cell.2016.11.039.

38. Yudovich, D., Bäckström, A., Schmiderer, L., Žemaitis, K., Subramaniam, A., and Larsson, J. (2020). Combined lentiviral- and RNA-mediated CRISPR/Cas9 delivery for efficient and traceable gene editing in human hematopoietic stem and progenitor cells. Sci. Rep. *10*, 22393. https://doi.org/10.1038/s41598-020-79724-x.

39. Nishimasu, H., Shi, X., Ishiguro, S., Gao, L., Hirano, S., Okazaki, S., Noda, T., Abudayyeh, O.O., Gootenberg, J.S., Mori, H., et al. (2018). Engineered CRISPR-Cas9 nuclease with expanded targeting space. Science *9*, 1259–1262. https://doi.org/10.1126/science.aas9129(2018).

40. Shifrut, E., Carnevale, J., Tobin, V., Roth, T.L., Woo, J.M., Bui, C.T., Li, P.J., Diolaiti, M.E., Ashworth, A., and Marson, A. (2018). Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. Cell *175*, 1958–1971.e15. https://doi.org/10.1016/j.cell.2018.10.024.

41. Ting, P.Y., Parker, A.E., Lee, J.S., Trussell, C., Sharif, O., Luna, F., Federe, G., Barnes, S.W., Walker, J.R., Vance, J., et al. (2018). Guide Swap enables genome-scale pooled CRISPR–Cas9 screening in human primary cells. Nat. Methods *15*, 941–946. https://doi.org/10.1038/s41592-018-0149-1.

42. Naldini, L. (2019). Genetic engineering of hematopoiesis: current stage of clinical translation and future perspectives. EMBO Mol. Med. *11*, e9958. https://doi.org/10.15252/emmm.201809958.

43. Kluesner, M.G., Lahr, W.S., Lonetree, C.-L., Smeester, B.A., Qiu, X., Slipek, N.J., Claudio Vázquez, P.N.C., Pitzen, S.P., Pomeroy, E.J., Vignes, M.J., et al. (2021). CRISPR-Cas9 cytidine and adenosine base editing of splice-sites mediates highly-efficient disruption of proteins in primary and immortalized cells. Nat. Commun. *12*, 2437. https://doi.org/10.1038/s41467-021-22009-2.

44. Borot, F., Wang, H., Ma, Y., Jafarov, T., Raza, A., Ali, A.M., and Mukherjee, S. (2019). Gene-edited stem cells enable CD33-directed immune therapy for myeloid malignancies. Proc. Natl. Acad. Sci. USA *116*, 11978–11987. https://doi.org/10.1073/pnas.1819992116.

45. Kim, M.Y., Yu, K.-R., Kenderian, S.S., Ruella, M., Chen, S., Shin, T.-H., Aljanahi, A.A., Schreeder, D., Klichinsky, M., Shestova, O., et al. (2018). Genetic inactivation of CD33 in hematopoietic stem cells to enable CAR T cell immunotherapy for acute myeloid leukemia. Cell *173*, 1439–1453.e19. https://doi.org/10.1016/j.cell.2018.05.013.

46. Humbert, O., Laszlo, G.S., Sichel, S., Ironside, C., Haworth, K.G., Bates, O.M., Beddoe, M.E., Carrillo, R.R., Kiem, H.-P., and Walter, R.B. (2019). Engineering resistance to CD33-targeted immunotherapy in normal hematopoiesis by CRISPR/Cas9-deletion of CD33 exon 2. Leukemia *33*, 762–808. https://doi.org/10.1038/s41375-018-0277-8.

47. Radtke, S., Adair, J.E., Giese, M.A., Chan, Y.-Y., Norgaard, Z.K., Enstrom, M., Haworth, K.G., Schefter, L.E., and Kiem, H.-P. (2017). A distinct hematopoietic stem cell population for rapid multilineage engraftment in nonhuman primates. Sci. Transl. Med. *9*, eaan1145. https://doi.org/10.1126/scitranslmed.aan1145.

48. McIntosh, B.E., Brown, M.E., Duffin, B.M., Maufort, J.P., Vereide, D.T., Slukvin, I.I., and Thomson, J.A. (2015). Nonirradiated NOD,B6.SCID Il2rγ-/- Kit(W41/W41) (NBSGW) mice support multilineage engraftment

of human hematopoietic cells. Stem Cell Rep. *4*, 171–180. https://doi.org/10.1016/j.stemcr.2014.12.005.

49. Martin-Rufino, J.D., and Sankaran, V.G. (2021). Deciphering transcriptional and functional heterogeneity in hematopoiesis with single-cell genomics. Curr. Opin. Hematol. *28*, 269–276. https://doi.org/10.1097/MOH.0000000000000657.

50. Replogle, J.M., Norman, T.M., Xu, A., Hussmann, J.A., Chen, J., Cogan, J.Z., Meer, E.J., Terry, J.M., Riordan, D.P., Srinivas, N., et al. (2020). Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. Nat. Biotechnol. *38*, 954–961. https://doi.org/10.1038/s41587-020-0470-y.

51. Cromer, M.K., Barsan, V.V., Jaeger, E., Wang, M., Hampton, J.P., Chen, F., Kennedy, D., Xiao, J., Khrebtukova, I., Granat, A., et al. (2022). Ultra-deep sequencing validates safety of CRISPR/Cas9 genome editing in human hematopoietic stem and progenitor cells. Nat. Commun. *13*, 4724. https://doi.org/10.1038/s41467-022-32233-z.

52. Sankaran, V.G., Menne, T.F., Xu, J., Akie, T.E., Lettre, G., Van Handel, B., Mikkola, H.K.A., Hirschhorn, J.N., Cantor, A.B., and Orkin, S.H. (2008). Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. Science *322*, 1839–1842. https://doi.org/10.1126/science.1165409.

53. Shen, Y., Verboon, J.M., Zhang, Y., Liu, N., Kim, Y.J., Marglous, S., Nandakumar, S.K., Voit, R.A., Fiorini, C., Ejaz, A., et al. (2021). A unified model of human hemoglobin switching through single-cell genome editing. Nat. Commun. *12*, 4991. https://doi.org/10.1038/s41467-021-25298-9.

54. Frangoul, H., Altshuler, D., Cappellini, M.D., Chen, Y.-S., Domm, J., Eustace, B.K., Foell, J., de la Fuente, J., Grupp, S., Handgretinger, R., et al. (2021). CRISPR-Cas9 gene editing for sickle cell disease and β-thalassemia. N. Engl. J. Med. *384*, 252–260. https://doi.org/10.1056/NEJMoa2031054.

55. Esrick, E.B., Lehmann, L.E., Biffi, A., Achebe, M., Brendel, C., Ciuculescu, M.F., Daley, H., MacKinnon, B., Morris, E., Federico, A., et al. (2021). Post-transcriptional genetic silencing of BCL11A to treat sickle cell disease. N. Engl. J. Med. *384*, 205–215. https://doi.org/10.1056/NEJMoa2029392.

56. Eisenstein, M. (2021). Gene Therapies Close in on a Cure for Sickle-Cell Disease (Nature Publishing Group) https://doi.org/10.1038/d41586-021-02138-w.

57. Ravi, N.S., Wienert, B., Wyman, S.K., Bell, H.W., George, A., Mahalingam, G., Vu, J.T., Prasad, K., Bandlamudi, B.P., Devaraju, N., et al. (2022). Identification of novel HPFH-like mutations by CRISPR base editing that elevate the expression of fetal hemoglobin. eLife *11*, e65421. https://doi.org/10.7554/eLife.65421.

58. Sankaran, V.G., and Orkin, S.H. (2013). The switch from fetal to adult hemoglobin. Cold Spring Harb. Perspect. Med. *3*, a011643. https://doi.org/10.1101/cshperspect.a011643.

59. Nandakumar, S.K., McFarland, S.K., Mateyka, L.M., Lareau, C.A., Ulirsch, J.C., Ludwig, L.S., Agarwal, G., Engreitz, J.M., Przychodzen, B., McConkey, M., et al. (2019). Gene-centric functional dissection of human genetic variation uncovers regulators of hematopoiesis. eLife *8*, e44080. https://doi.org/10.7554/eLife.44080.

60. Ulirsch, J.C., Lacy, J.N., An, X., Mohandas, N., Mikkelsen, T.S., and Sankaran, V.G. (2014). Altered chromatin occupancy of master regulators underlies evolutionary divergence in the transcriptional landscape of erythroid differentiation. PLoS Genet. *10*, e1004890. https://doi.org/10.1371/journal.pgen.1004890.

61. Nam, A.S., Dusaj, N., Izzo, F., Murali, R., Myers, R.M., Mouhieddine, T.H., Sotelo, J., Benbarche, S., Waarts, M., Gaiti, F., et al. (2022). Single-cell multi-omics of human clonal hematopoiesis reveals that DNMT3A R882 mutations perturb early progenitor states through selective hypomethylation. Nat. Genet. *54*, 1514–1526. https://doi.org/10.1038/s41588-022-01179-9.

62. Beneyto-Calabuig, S., Ludwig, A.K., Kniffka, J.-A., Szu-Tu, C., Rohde, C., Antes, M., Waclawiczek, A., Gräßle, S., Pervan, P., Janssen, M., et al. (2022). Clonally resolved single-cell multi-omics identifies routes of cellular differentiation in acute myeloid leukemia. Preprint at bioRxiv. https://doi.org/10.1101/2022.08.29.505648.

63. Wienert, B., Funnell, A.P.W., Norton, L.J., Pearson, R.C.M., Wilkinson-White, L.E., Lester, K., Vadolas, J., Porteus, M.H., Matthews, J.M., Quinlan, K.G.R., et al. (2015). Editing the genome to introduce a beneficial naturally occurring mutation associated with increased fetal globin. Nat. Commun. *6*, 7085. https://doi.org/10.1038/ncomms8085.

64. Wienert, B., Martyn, G.E., Funnell, A.P.W., and Quinlan, K.G.R. (2018). Wake-up sleepy gene: reactivating fetal globin for b -Hemoglobinopathies. Trends Genet. *34*, 927–940. https://doi.org/10.1016/j.tig.2018.09.004.

65. Antoniou, P., Hardouin, G., Martinucci, P., Frati, G., Felix, T., Chalumeau, A., Fontana, L., Martin, J., Masson, C., Brusson, M., et al. (2022). Base-editing-mediated dissection of a γ-globin cis-regulatory element for the therapeutic reactivation of fetal hemoglobin expression. Nat. Commun. *13*, 6618. https://doi.org/10.1038/s41467-022-34493-1.

66. Doerfler, P.A., Feng, R., Li, Y., Palmer, L.E., Porter, S.N., Bell, H.W., Crossley, M., Pruett-Miller, S.M., Cheng, Y., and Weiss, M.J. (2021). Activation of γ-globin gene expression by GATA1 and NF-Y in hereditary persistence of fetal hemoglobin. Nat. Genet. *53*, 1177–1186. https://doi.org/10.1038/s41588-021-00904-0.

67. Martyn, G.E., Wienert, B., Yang, L., Shah, M., Norton, L.J., Burdach, J., Kurita, R., Nakamura, Y., Pearson, R.C.M., Funnell, A.P.W., et al. (2018). Natural regulatory mutations elevate the fetal globin gene via disruption of BCL11A or ZBTB7A binding. Nat. Genet. *50*, 498–503. https://doi.org/10.1038/s41588-018-0085-0.

68. Liu, N., Hargreaves, V.V., Zhu, Q., Kurland, J.V., Hong, J., Kim, W., Sher, F., Macias-Trevino, C., Rogers, J.M., Kurita, R., et al. (2018). Direct Promoter Repression by BCL11A Controls the Fetal to Adult Hemoglobin Switch. Cell *173*, 430–442.e17. https://doi.org/10.1016/j.cell.2018.03.016.

69. Crispino, J.D., and Horwitz, M.S. (2017). GATA factor mutations in hematologic disease. Blood *129*, 2103–2110. https://doi.org/10.1182/BLOOD-2016-09-687889.

70. Sankaran, V.G., Ghazvinian, R., Do, R., Thiru, P., Vergilio, J.A., Beggs, A.H., Sieff, C.A., Orkin, S.H., Nathan, D.G., Lander, E.S., et al. (2012). Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. J. Clin. Invest. *122*, 2439–2443. https://doi.org/10.1172/JCI63597.

71. Nichols, K.E., Crispino, J.D., Poncz, M., White, J.G., Orkin, S.H., Maris, J.M., and Weiss, M.J. (2000). Familial dyserythropoietic anaemia and thrombocytopenia due to an inherited mutation in GATA1. Nat. Genet. *24*, 266–270. https://doi.org/10.1038/73480.

72. Phillips, J.D., Steensma, D.P., Pulsipher, M.A., Spangrude, G.J., and Kushner, J.P. (2007). Congenital erythropoietic porphyria due to a mutation in GATA1: the first trans-acting mutation causative for a human porphyria. Blood *109*, 2618–2621. https://doi.org/10.1182/blood-2006-06-022848.

73. Yu, C., Niakan, K.K., Matsushita, M., Stamatoyannopoulos, G., Orkin, S.H., and Raskind, W.H. (2002). X-linked thrombocytopenia with thalassemia from a mutation in the amino finger of GATA-1 affecting DNA binding rather than FOG-1 interaction. Blood *100*, 2040–2045. https://doi.org/10.1182/blood-2002-02-0387.

74. Tubman, V.N., Levine, J.E., Campagna, D.R., Monahan-Earley, R., Dvorak, A.M., Neufeld, E.J., and Fleming, M.D. (2007). X-linked gray platelet syndrome due to a GATA1 Arg216Gln mutation. Blood *109*, 3297–3299. https://doi.org/10.1182/blood-2006-02-004101.

75. Hasle, H., Kline, R.M., Kjeldsen, E., Nik-Abdul-Rashid, N.F., Bhojwani, D., Verboon, J.M., DiTroia, S.P., Chao, K.R., Raaschou-Jensen, K., Palle, J., et al. (2022). Germline GATA1s-generating mutations predispose to

leukemia with acquired trisomy 21 and Down syndrome-like phenotype. Blood *139*, 3159–3165. https://doi.org/10.1182/blood.2021011463.

76. Wechsler, J., Greene, M., McDevitt, M.A., Anastasi, J., Karp, J.E., Le Beau, M.M., and Crispino, J.D. (2002). Acquired mutations in GATA1 in the megakaryoblastic leukemia of Down syndrome. Nat. Genet. *32*, 148–152. https://doi.org/10.1038/ng955.

77. Ludwig, L.S., Lareau, C.A., Bao, E.L., Liu, N., Utsugisawa, T., Tseng, A.M., Myers, S.A., Verboon, J.M., Ulirsch, J.C., Luo, W., et al. (2022). Congenital anemia reveals distinct targeting mechanisms for master transcription factor GATA1. Blood *139*, 2534–2546. https://doi.org/10.1182/blood.2021013753.

78. Qiu, X., Zhang, Y., Martin-Rufino, J.D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A.N., Hein, M.Y., Hoi Joseph Min, K., Wang, L., et al. (2022). Mapping transcriptomic vector fields of single cells. Cell *185*, 690–711.e45. https://doi.org/10.1016/j.cell.2021.12.045.

79. Voit, R.A., Tao, L., Yu, F., Cato, L.D., Cohen, B., Fleming, T.J., Antoszewski, M., Liao, X., Fiorini, C., Nandakumar, S.K., et al. (2023). A genetic disorder reveals a hematopoietic stem cell regulatory network co-opted in leukemia. Nat. Immunol. *24*, 69–83. https://doi.org/10.1038/s41590-022-01370-4.

80. Kratz, C.P., Niemeyer, C.M., Karow, A., Volz-Fleckenstein, M., Schmitt-Gräff, A., and Strahm, B. (2008). Congenital transfusion-dependent anemia and thrombocytopenia with myelodysplasia due to a recurrent GATA1G208R germline mutation. Leukemia *22*, 432–434. https://doi.org/10.1038/sj.leu.2404904.

81. Freson, K., Matthijs, G., Thys, C., Mariën, P., Hoylaerts, M.F., Vermylen, J., and Van Geet, C. (2002). Different substitutions at residue D218 of the X-linked transcription factor GATA1 lead to altered clinical severity of macrothrombocytopenia and anemia and are associated with variable skewed X inactivation. Hum. Mol. Genet. *11*, 147–152. https://doi.org/10.1093/hmg/11.2.147.

82. Bastida, J.M., Malvestiti, S., Boeckelmann, D., Palma-Barqueros, V., Wolter, M., Lozano, M.L., Glonnegger, H., Benito, R., Zaninetti, C., Sobotta, F., et al. (2022). A novel GATA1 variant in the C-terminal zinc finger compared with the platelet phenotype of patients with A likely pathogenic variant in the N-terminal zinc finger. Cells *11*, 3223. https://doi.org/10.3390/cells11203223.

83. Khajuria, R.K., Munschauer, M., Ulirsch, J.C., Fiorini, C., Ludwig, L.S., McFarland, S.K., Abdulhay, N.J., Specht, H., Keshishian, H., Mani, D.R., et al. (2018). Ribosome levels selectively regulate translation and lineage commitment in human hematopoiesis. Cell *173*, 90–103.e19. https://doi.org/10.1016/J.CELL.2018.02.036.

84. Papalexi, E., Mimitou, E.P., Butler, A.W., Foster, S., Bracken, B., Mauck, W.M., Wessels, H.-H., Hao, Y., Yeung, B.Z., Smibert, P., et al. (2021). Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. Nat. Genet. *53*, 322–331. https://doi.org/10.1038/s41588-021-00778-2.

85. Burke, W., Parens, E., Chung, W.K., Berger, S.M., and Appelbaum, P.S. (2022). The challenge of genetic variants of uncertain clinical significance: A narrative review. Ann. Intern. Med. *175*, 994–1000. https://doi.org/10.7326/M21-4109.

86. Thuronyi, B.W., Koblan, L.W., Levy, J.M., Yeh, W.-H., Zheng, C., Newby, G.A., Wilson, C., Bhaumik, M., Shubina-Oleinik, O., Holt, J.R., et al. (2019). Continuous evolution of base editors with expanded target compatibility and improved activity. Nat. Biotechnol. *37*, 1070–1079. https://doi.org/10.1038/s41587-019-0193-0.

87. Jarocha, D., Vo, K.K., Lyde, R.B., Hayes, V., Camire, R.M., and Poncz, M. (2018). Enhancing functional platelet release in vivo from in vitro-grown megakaryocytes using small molecule inhibitors. Blood Adv. *2*, 597–606. https://doi.org/10.1182/bloodadvances.2017010975.

88. Guo, M.H., Nandakumar, S.K., Ulirsch, J.C., Zekavat, S.M., Buenrostro, J.D., Natarajan, P., Salem, R.M., Chiarle, R., Mitt, M., Kals, M., et al. (2017). Comprehensive population-based genome sequencing provides insight into hematopoietic regulatory mechanisms. Proc. Natl. Acad. Sci. USA *114*, E327–E336. https://doi.org/10.1073/pnas.1619052114.

89. Smith, S.L., Bender, J.G., Maples, P.B., Unverzagt, K., Schilling, M., Lum, L., Williams, S., and Van Epps, D.E. (1993). Expansion of neutrophil precursors and progenitors in suspension cultures of CD34+ cells enriched from human bone marrow. Exp. Hematol. *21*, 870–877.

90. Stec, M., Weglarczyk, K., Baran, J., Zuba, E., Mytar, B., Pryjma, J., and Zembala, M. (2007). Expansion and differentiation of CD14+CD16(-) and CD14+ +CD16+ human monocyte subsets from cord blood CD34+ hematopoietic progenitors. J. Leukoc. Biol. *82*, 594–602. https://doi.org/10.1189/jlb.0207117.

91. Lee, J., Breton, G., Oliveira, T.Y.K., Zhou, Y.J., Aljoufi, A., Puhr, S., Cameron, M.J., Sékaly, R.P., Nussenzweig, M.C., and Liu, K. (2015). Restricted dendritic cell and monocyte progenitors in human cord blood and bone marrow. J. Exp. Med. *212* (3), 385–399. https://doi.org/10.1084/jem.20141442.

92. Hernández, D.C., Juelke, K., Müller, N.C., Durek, P., Ugursu, B., Mashreghi, M.-F., Rückert, T., and Romagnani, C. (2021). An in vitro platform supports generation of human innate lymphoid cells from CD34+ hematopoietic progenitors that recapitulate ex vivo identity. Immunity *54*, 2417–2432.e5. https://doi.org/10.1016/j.immuni.2021.07.019.

93. Mrózek, E., Anderson, P., and Caligiuri, M.A. (1996). Role of interleukin-15 in the development of human CD56+ natural killer cells from CD34+ hematopoietic progenitor cells. Blood *87*, 2632–2640.

94. Seet, C.S., He, C., Bethune, M.T., Li, S., Chick, B., Gschweng, E.H., Zhu, Y., Kim, K., Kohn, D.B., Baltimore, D., et al. (2017). Generation of mature T cells from human hematopoietic stem and progenitor cells in artificial thymic organoids. Nat. Methods *14*, 521–530. https://doi.org/10.1038/nmeth.4237.

95. Luo, X.M., Maarschalk, E., O'Connell, R.M., Wang, P., Yang, L., and Baltimore, D. (2009). Engineering human hematopoietic stem/progenitor cells to produce a broadly neutralizing anti-HIV antibody after in vitro maturation to human B lymphocytes. Blood *113*, 1422–1431. https://doi.org/10.1182/blood-2008-09-177139.

96. Ursu, O., Neal, J.T., Shea, E., Thakore, P.I., Jerby-Arnon, L., Nguyen, L., Dionne, D., Diaz, C., Bauman, J., Mosaad, M.M., et al. (2022). Massively parallel phenotyping of coding variants in cancer with Perturb-seq. Nat. Biotechnol. *40*, 896–905. https://doi.org/10.1038/s41587-021-01160-7.

97. Schaid, D.J., Chen, W., and Larson, N.B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat. Rev. Genet. *19*, 491–504. https://doi.org/10.1038/S41576-018-0016-Z.

98. Huang, T.P., Heins, Z.J., Miller, S.M., Wong, B.G., Balivada, P.A., Wang, T., Khalil, A.S., and Liu, D.R. (2022). High-throughput continuous evolution of compact Cas9 variants targeting single-nucleotide-pyrimidine PAMs. Nat. Biotechnol. *41*, 96–107. https://doi.org/10.1038/s41587-022-01410-2.

99. Walton, R.T., Christie, K.A., Whittaker, M.N., and Kleinstiver, B.P. (2020). Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. Science *368*, 290–296. https://doi.org/10.1126/science.aba8853.

100. Anzalone, A.V., Randolph, P.B., Davis, J.R., Sousa, A.A., Koblan, L.W., Levy, J.M., Chen, P.J., Wilson, C., Newby, G.A., Raguram, A., et al. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. Nature *576*, 149–157. https://doi.org/10.1038/s41586-019-1711-4.

101. Nuñez, J.K., Chen, J., Pommier, G.C., Cogan, J.Z., Replogle, J.M., Adriaens, C., Ramadoss, G.N., Shi, Q., Hung, K.L., Samelson, A.J., et al. (2021). Genome-wide programmable transcriptional memory by CRISPR-based epigenome editing. Cell *184*, 2503–2519.e17. https://doi.org/10.1016/j.cell.2021.03.025.

102. Amabile, A., Migliara, A., Capasso, P., Biffi, M., Cittaro, D., Naldini, L., and Lombardo, A. (2016). Inheritable silencing of endogenous genes by hit-and-run targeted epigenetic editing. Cell *167*, 219–232.e14. https://doi.org/10.1016/j.cell.2016.09.006.

103. Petri, K., Zhang, W., Ma, J., Schmidts, A., Lee, H., Horng, J.E., Kim, D.Y., Kurt, I.C., Clement, K., Hsu, J.Y., et al. (2022). CRISPR prime editing with ribonucleoprotein complexes in zebrafish and primary human cells. Nat. Biotechnol. *40*, 189–193. https://doi.org/10.1038/s41587-021-00901-y.

104. Koblan, L.W., Arbab, M., Shen, M.W., Hussmann, J.A., Anzalone, A.V., Doman, J.L., Newby, G.A., Yang, D., Mok, B., Replogle, J.M., et al. (2021). Efficient C• G-to-G• C base editors developed using CRISPRi screens, target-library analysis, and machine learning. Nat. Biotechnol. *39*, 1414–1425.

105. Newby, G.A., Yen, J.S., Woodard, K.J., Mayuranathan, T., Lazzarotto, C.R., Li, Y., Sheppard-Tillman, H., Porter, S.N., Yao, Y., Mayberry, K., et al. (2021). Base editing of haematopoietic stem cells rescues sickle cell disease in mice. Nature *595*, 295–302. https://doi.org/10.1038/s41586-021-03609-w.

106. Naeem, M., Majeed, S., Hoque, M.Z., and Ahmad, I. (2020). Latest developed strategies to minimize the off-target effects in CRISPR-Cas-mediated genome editing. Cells *9*, 1608. https://doi.org/10.3390/cells9071608.

107. Carroll, D. (2019). Collateral damage: benchmarking off-target effects in genome editing. Genome Biol. *20*, 114. https://doi.org/10.1186/s13059-019-1725-0.

108. Rodriguez-Meira, A., Buck, G., Clark, S.-A., Povinelli, B.J., Alcolea, V., Louka, E., McGowan, S., Hamblin, A., Sousos, N., Barkas, N., et al. (2019). Unravelling intratumoral heterogeneity through high-sensitivity single-cell mutational analysis and parallel RNA sequencing. Mol. Cell *73*, 1292–1305.e8. https://doi.org/10.1016/j.molcel.2019.01.009.

109. Giani, F.C., Fiorini, C., Wakabayashi, A., Ludwig, L.S., Salem, R.M., Jobaliya, C.D., Regan, S.N., Ulirsch, J.C., Liang, G., Steinberg-Shemer, O., et al. (2016). Targeted application of human genetic variation can improve red blood cell production from stem cells. Cell Stem Cell *18*, 73–78. https://doi.org/10.1016/j.stem.2015.09.015.

110. Sanjana, N.E., Shalem, O., and Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. Nat. Methods *11*, 783–784. https://doi.org/10.1038/nmeth.3047.

111. Hill, A.J., McFaline-Figueroa, J.L., Starita, L.M., Gasperini, M.J., Matreyek, K.A., Packer, J., Jackson, D., Shendure, J., and Trapnell, C. (2018). On the design of CRISPR-based single-cell molecular screens. Nat. Methods *15*, 271–274. https://doi.org/10.1038/nmeth.4604.

112. Sanson, K.R., Hanna, R.E., Hegde, M., Donovan, K.F., Strand, C., Sullender, M.E., Vaimberg, E.W., Goodale, A., Root, D.E., Piccioni, F., et al. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nat. Commun. *9*, 5416. https://doi.org/10.1038/s41467-018-07901-8.

113. Dang, Y., Jia, G., Choi, J., Ma, H., Anaya, E., Ye, C., Shankar, P., and Wu, H. (2015). Optimizing sgRNA structure to improve CRISPR-Cas9 knockout efficiency. Genome Biol. *16*, 280. https://doi.org/10.1186/s13059-015-0846-3.

114. Gutierrez, C., Al'Khafaji, A.M., Brenner, E., Johnson, K.E., Gohil, S.H., Lin, Z., Knisbacher, B.A., Durrett, R.E., Li, S., Parvin, S., et al. (2021). Multifunctional barcoding with ClonMapper enables high-resolution study of clonal dynamics during tumor evolution and treatment. Nat. Cancer *2*, 758–772. https://doi.org/10.1038/s43018-021-00222-8.

115. Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T.L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics *13*, 134. https://doi.org/10.1186/1471-2105-13-134.

116. Al'Khafaji, A.M., Smith, J.T., Garimella, K.V., Babadi, M., Sade-Feldman, M., Gatzen, M., Sarkizova, S., Schwartz, M.A., Popic, V., Blaum, E.M., et al. (2021). High-throughput RNA isoform sequencing using programmable cDNA concatenation. Preprint at bioRxiv. https://doi.org/10.1101/2021.10.01.462818.

117. Clement, K., Rees, H., Canver, M.C., Gehrke, J.M., Farouni, R., Hsu, J.Y., Cole, M.A., Liu, D.R., Joung, J.K., Bauer, D.E., et al. (2019). CRISPResso2 provides accurate and rapid genome editing sequence analysis. Nat. Biotechnol. *37*, 224–226. https://doi.org/10.1038/s41587-019-0032-3.

118. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419. https://doi.org/10.1038/nmeth.4197.

119. Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15. https://doi.org/10.1186/s13059-017-1382-0.

120. Bae, S., Park, J., and Kim, J.-S. (2014). Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. Bioinformatics *30*, 1473–1475. https://doi.org/10.1093/bioinformatics/btu048.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CD3 BV605 Anti-human [Clone: SK7]	BioLegend	Cat: 344836; RRID: AB_2565825
CD11b Pacific Blue Anti-human [Clone: IICRF44]	BioLegend	Cat: 301315; RRID: AB_493015
CD19 FITC Anti-human [Clone: REA675]	Miltenyi Biotec	Cat: 130-113-645; RRID: AB_2726198
CD33 PE Anti-human [Clone: HIM3-4]	ThermoFisher	Cat: 12-0339-42; RRID: AB_10855031
CD33 APC Anti-human [Clone: WM53]	BioLegend	Cat: 983902; RRID: AB_2810824
CD33 FITC Anti-human [Clone: HIM3-4]	Invitrogen	Cat: 11-0339-42; RRID: AB_10718827
CD34 BV421 Anti-human [Clone: 561]	BioLegend	Cat: 343610; RRID: AB_2561358
CD45 PE Anti-human [Clone: 2D1]	BioLegend	Cat: 368510; RRID: AB_2566370
CD45 APC Anti-human [Clone: HI30]	BioLegend	Cat: 304037; RRID: AB_2562049
CD45 FITC Anti-mouse [Clone: 30-F11]	BioLegend	Cat: 103108; RRID: AB_312973
CD45 APC Anti-mouse [Clone: 30-F11]	BioLegend	Cat: 103112; RRID: AB_312977
CD71 APC Anti-human [Clone: OKT9]	eBioscience	Cat: 17-0719-42; RRID: AB_10671393
CD123 PE-Cy7 Anti-human [Clone: 6H6]	BioLegend	Cat: 306010; RRID: AB_493576
Brilliant Violet 421™ anti-human CD71 [Clone: CY1G4]	BioLegend	Cat: 334122; RRID: AB_2734337
APC anti-human CD235a (Glycophorin A) [Clone: HI264]	BioLegend	Cat: 349114; RRID: AB_2650976
Fetal Hemoglobin APC Monoclonal Antibody [Clone: HBF-1]	Life Technologies	Cat: MHFH05; RRID: AB_10374595
CD235a PE Anti-human [Clone: HIR2]	Invitrogen	Cat: 12-9987-82; RRID: AB_466300
7-AAD Viability Staining Solution	BioLegend	Cat: 420404
7-AAD	BD Biosciences	Cat: 51-68981E
Bacterial and Virus Strains		
BL21 Star DE3 Competent Cells	ThermoFisher	Cat: C601003
Endura™ Electrocompetent Cells	Biosearch Technologies	Cat: 71003-038
OneShot TOP10 Chemically Competent Cells	Invitrogen	Cat: C404006
Biological samples		
Human CD34+ hematopoietic stem and progenitor cells, adult	Fred Hutchinson Cancer Research Center	N/A
Cord Blood Unit for umbilical cord-derived CD34+ hematopoietic stem and progenitor cells	Dana Farber Pasquarello Tissue Bank	N/A
Chemicals, peptides, and recombinant proteins		
Kanamycin	Sigma-Aldrich	Cat: K4000-25G
L-Rhamnose	Sigma-Aldrich	Cat: R3875-100G
HEPES 7.5 pH	Gibco	Cat: 15630080
TCEP	Thomas Scientific	Cat: 51805-45-9
Roche EDTA-free complete protease inhibitor cocktail	Roche	Cat: 11697498001
DNase I Solution	Thermo Scientific	Cat: 90083
Imidazole	GoldBio	Cat: I-901-25
Dulbecco's Modified Eagle Medium-High Glucose (DMEM)	Life Technologies	Cat: 11965-118

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Iscove's Modified Dulbecco's Medium (IMDM)	Life Technologies	Cat: 12440-061
Human Holo-Transferrin	Sigma-Aldrich	Cat: T0665
Fetal Bovine Serum (FBS)	BioTechne	Cat: S11550
Penicillin-Streptomycin	GIBCO	Cat: 15140-122
Human Serum, Type AB	Atlanta Biologicals	Cat: S40110
Human Plasma, Type AB	SeraCare	Cat: 1810-0001
Humulin R (Insulin)	Lilly	Cat: NDC 0002-8215-01
Heparin	Hospira	Cat: NDC 00409-2720-01
Epogen (recombinant erythropoietin)	Amgen	Cat: NDC 55513-267-10
Recombinant human stem cell factor (SCF)	PeproTech	Cat: 300-07
Recombinant human interleukin-3 (IL3)	PeproTech	Cat: 200-03
Opti-MEM	GIBCO	Cat: 31985-062
StemSpan SFEM II medium	StemCell Technologies	Cat: 02690
StemSpan CC100	StemCell Technologies	Cat: 02690
FuGENE 6 Transfection Reagent	Promega	Cat: E2691
Polybrene Infection/Transfection reagent	Millipore	Cat: TR-1003-G
Recombinant Human Thrombopoietin	PeproTech	Cat: 300-18
UM171	StemCell Technologies	Cat: 72912
2-mercaptoethanol	Sigma	Cat: M6250
PBS	GIBCO	Cat: 10010-023
1X SPRI beads	Beckman Coulter Inc	Cat: B23318
L-Glutamine	Thermo Fisher Scientific	Cat: 25-030-081
Penicillin/Streptomycin	Life Technologies	Cat: 15140-122
FastDigest Esp3I	Thermo Scientific	Cat: FD0454
SYBR green 10X	VWR International LLC	Cat: 12001-796
MluI	Thermo Fisher Scientific	Cat: FERFD0564
PspLI	Thermo Fisher Scientific	Cat: FERFD0854
T4 DNA Ligase Reaction Buffer	New England Biolabs	Cat: B0202S
T7 Ligase	Qiagen Beverly LLC	Cat: L6020L
10% Glutaraldehyde	Electron Microscopy Sciences	Cat: 16121
Triton X-100	Sigma-Aldrich	Cat: X100-1L
Bovine Serum Albumin	Sigma-Aldrich	Cat: A9418
Critical commercial assays		
MinElute PCR Purification Kit	QIAGEN	Cat: 28004
NEBNext® High-Fidelity 2X PCR Master Mix	New England Biolabs	Cat: M0541L
KAPA HiFi HotStart PCR ReadyMix	KAPA HiFi HotStart PCR ReadyMix	Cat: KK2602
Qubit dsDNA HS Assay Kit	Thermo Fisher	Cat: Q32854
Bioanalyzer High Sensitivity DNA Analysis	Agilent	Cat: 5067-4626
QIAquick Gel Extraction Kit	Qiagen	Cat: 28706X4
NucleoBond Xtra Maxi	Macherey-Nagel	Cat: 740424.50
QIAamp DNA Micro Kit (50)	Qiagen	Cat: 56304
QIAamp DNA Mini Kit (50)	Qiagen	Cat: 51304
QIAquick PCR Purification Kit	Qiagen	Cat: 28104
Q5 Hot Start High-Fidelity 2X Master Mix	New England Biolabs	Cat: M0492L
EasySep™ Human Cord Blood CD34 Positive Selection Kit II	StemCell Technologies	Cat: 17896
GE Healthcare HiTrap SP HP	Cytiva	17115101

(*Continued on next page*)

Continued		
REAGENT or RESOURCE	SOURCE	IDENTIFIER
NuPAGE 4 to 12% Bis-Tris gel	Invitrogen	NP0321BOX
Amicon Ultra-0.5 Centrifugal Filter Units	Sigma-Aldrich	UFC5100BK
Quick-DNA- FFPE kit	Zymo	D3067
Deposited data		
Perturb(BE)-seq, scRNA-seq and bulk RNA-seq	This study	GEO: GSE215253
Editing efficiency amplicon sequencing, pooled single-cell genotyping, and functional screening sgRNA sequencing	This study	PRJNA889818
Experimental models: Cell lines		
293T cells	ATCC	Cat: CRL-3216
Experimental models: Organisms/strains		
Mouse: NOD.Cg-*Kit*$^{W\text{-}41J}$*Tyr*$^{+}$*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/ThomJ (NBSGW)	Jackson Laboratory	IMSR_JAX:026622
Oligonucleotides		
Primers, oligonucleotides and sgRNAs	IDT	Tables S1, S2, S3, S4, and S5
Recombinant DNA		
Modified CROP-seq vector for hematopoiesis screens	This study	N/A
Protein purification vector: bacterial codon optimized ABE8e-Cas9NG	This study	N/A
Protein purification vector: bacterial codon optimized evoFERNY-Cas9NG	This study	N/A
Software and algorithms		
Original code	This study	https://github.com/sankaranlab/hematopoiesis_be_screens https://doi.org/10.5281/zenodo.7781053
Base editor sgRNA design tool	Github	https://github.com/mhegde/base-editor-design-tool
BCL Convert v4.0.3	Illumina	N/A
Bcl2fastq2 v2.20	Illumina	N/A
CRISPResso2 v2.0.45	Clement et al.[117]	https://doi.org/10.1038/s41587-019-0032-3
CellRanger v6.0.1	10X Genomics	N/A
Tapestri Pipeline v2	MissionBio	N/A
PHAST package	Hubisz et al.	http://compgen.bscb.cornell.edu/phast/
The PyMOL Molecular Graphics System, Version 2.0	Schrödinger, LLC.	https://pymol.org/
Salmon v1.10.0	Patro et al.[118]	https://combine-lab.github.io/salmon/
kb-python	Melsted et al.	https://github.com/pachterlab/kb_python
Dynamo	Qiu et al.	https://github.com/aristoteleo/dynamo-release

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Vijay G. Sankaran (sankaran@broadinstitute.org).

Materials availability

All plasmids described are available upon request to the lead contact.

Data and code availability

- Single-cell RNA-seq data have been deposited at GEO and are publicly available as of the date of publication. Single-cell genotyping, editing efficiency and sequence data have been deposited at SRA and are all publicly available as of the date of publication. Accession numbers are listed in the key resources table.
- All original code has been deposited at Github and Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Primary human cells

Primary human CD34+ hematopoietic stem and progenitor cells (HSPCs) from mobilized peripheral blood of healthy donors were purchased from the Fred Hutchinson Cancer Research Center.

Mice

For xenotransplantation assays NOD.Cg-*Kit*$^{W\text{-}41J}$*Tyr*$^{+}$*Prkdc*^{scid}*Il2rg*^{tm1Wjl}/ThomJ (NBSGW) mice were obtained from Jackson Laboratory (026622).[48] Littermates of the same sex (4 males and 5 females) were randomly assigned to experimental groups, and were 8 weeks old at the time of transplantation. All animal experiments were approved by the Boston Children's Hospital Institutional Animal Care and Use Committee (A3303-01).

METHOD DETAILS

Base editor protein plasmids

Bacterial protein purification plasmids expressing ABE8e-Cas9NG (https://www.addgene.org/138489/) or evoFERNY (https://www.addgene.org/125615/) Cas9NG, were bacterial codon-optimized using Genscript. Fragments were ordered as gBlocks (IDT) and assembled using USER cloning into an N-terminal His-tag inducible bacterial expression plasmid.

Base editor protein purification

We optimized previously described base editor protein purification protocols.[7] Briefly, the base editor protein expression plasmids were transformed into BL21 Star DE3 competent cells (Thermo Fisher, C601003). Bacteria were inoculated in Terrific Broth (TB) supplemented with 25 μg/mL kanamycin (Sigma-Aldrich, K4000-25G) at 37°C until the optical density at 600 nm reached 1.5. The culture was then cooled down and supplemented with 30% L-rhamnose (Sigma-Aldrich, R3875-100G) to a final concentration of 0.8% to induce protein expression at 18°C for 24 hours. The cell pellets were collected and flash frozen and stored at -80°C. The procedures that follow were performed on the same day and at 4°C. Cell pellets from 1L of culture were resuspended in 30 mL cold bacterial lysis buffer (20 mM HEPES 7.5 (Gibco, 15630080), 2 M NaCl, 10% Glycerol, 1 mM TCEP (Thomas Scientific, 51805-45-9), 2 tablets of Roche EDTA-free complete protease inhibitor cocktail (Roche, 11697498001), 75 U DNase I solution (Thermo Scientific, 90083) followed by lysis twice by a homogenizer (Microfluidics) at ~18,000 psi. The lysate was centrifuged at 40,000 g at 4°C for 30 minutes. The supernatant was collected and incubated with 0.75 mL Ni-Penta resins (Marvelgent Bioscience, 11-0227-010) at 4°C for 1 hour. Subsequently, the solution was flowed through a disposable chromatography column (QIAGEN, 30210) at 4°C. The column was further washed with 100 mL wash buffer (20 mM HEPES 7.5, 2 M NaCl, 10% Glycerol, 1 mM TCEP, 25 mM Imidazole (GoldBio, I-901-25). The protein was subsequently eluted with an elution buffer (20 mM HEPES 7.5, 10% Glycerol, 1 mM TCEP, 500 mM Imidazole). The elution was collected and analyzed by SDS-PAGE before a further purification step by cation exchange chromatography using GE Healthcare HiTrap SP HP (Cytiva, 17115101) on an Äkta Pure25 FPLC system (Cytiva, 29018224). The low salt buffer for cation exchange chromatography was prepared using 20 mM HEPES 7.5, 10% Glycerol, and 1 mM TCEP. The high salt buffer was prepared using 20 mM HEPES 7.5, 2 M NaCl, 10% Glycerol, and 1 mM TCEP. The purification fractions were run on an NuPAGE 4 to 12% Bis-Tris gel (Invitrogen, NP0321BOX), and fractions of the right size were pooled and concentrated with a Amicon Ultra-0.5 Centrifugal Filter Unit (Sigma, UFC5100BK) to around 60 μM (determined by Nanodrop) and flash frozen and stored at -80°C.

Cell culture

Primary human CD34+ hematopoietic stem and progenitor cells (HSPCs) from mobilized peripheral blood of healthy donors were purchased from the Fred Hutchinson Cancer Research Center. Cells were cultured in StemSpan SFEM II human hematopoietic stem cell expansion media (StemCell Technologies, 02690) with 1% L-glutamine (Thermo Fisher Scientific, 25-030-081) and 1% penicillin/streptomycin (Life Technologies, 15140-122) and supplemented with the CC100 cytokine cocktail 100x (StemCell

Technologies, 02690) and 50 ng/mL human TPO (PeproTech, 300-18). This strategy allows for expansion of CD34+ cells on the initial days of culture and multilineage hematopoietic differentiation over the course of 7–10 days.

In the benchmarking experiments (Figures 1, S1, and S7), as well as the CD33 base editor screens in HSPCs, the transplant into NBSGW mice (Figures 2 and 7), and the *GATA1* pooled single-cell genotyping experiments (Figure 5) the stem cell expansion culture media described above was additionally supplemented with 35 μM UM171 (Stem Cell Technologies, 72912) to maximize maintenance of undifferentiated HSPCs.

To induce erythroid differentiation (Figures 4, 5, 6, and 7), HSPCs were differentiated into red blood cells utilizing a three-phase culture protocol, as we have described and characterized previously.[59,109] In phase 1 (day 0–7), cells were cultured in IMDM (Life Technologies, 12440-061) supplemented with 3% human AB serum (Atlanta Biologicals, S40110), 2% human AB plasma (SeraCare, 1810-0001), 1% penicillin/streptomycin (Life Technologies, 15140-122), 10 μg/mL insulin (Lilly, NDC 0002-8215-01), 3 IU/mL heparin (Hospira, NDC 00409-2720-01), 200 μg/mL holo-transferrin (Sigma-Aldrich, T0665), 10 ng/mL stem cell factor (SCF) (Peprotech, 300-07), 1 IU erythropoietin (EPO) (Amgen, NDC 55513-267-10) and 1 ng/mL IL-3 (Peprotech, 200-03). In phase 2 (days 8–13), IL-3 was omitted from the medium. In phase 3 (day 14–21), both IL-3 and SCF were omitted from the medium and the holo-transferrin concentration was increased to 1 mg/mL. In the fetal hemoglobin functional screen, cells were profiled on day 13 of erythroid differentiation. For colony-forming assays, HSPCs were plated in MethoCult H4434 (StemCell Technologies, 04434), in three 35mm dishes (technical replicates) for each biological replicate and grown at 37°C with 5% CO_2 for 14 days.

In the *HBG1/2* promoter and *GATA1* systematic mutagenesis screens (Figures 4, 5, and 6), cells were initially transduced and cultured in the expansion medium described above for 4 days, transitioned to phase I erythroid media, and electroporated on day 5. For the *HBG1/2* promoter screen, cells were profiled on day 13 of erythroid differentiation, prior to enucleation. In the experimental validation of the GATA1 VUSes (Figures 6 and 7) cells were cultured in expansion media for 2-3 days prior to electroporation.

Base editor protein electroporation

Cells were washed three times in DPBS prior to electroporation, to remove any residual RNAses present in the culture media that could interfere with editing. We first diluted freshly thawed base editor protein in P3 Lonza buffer with supplement (Lonza, V4XP-3032). The final amount of protein per electroporation ranged between 20-40 μg, and was optimized using the base editing activity of the batch as assessed on titration experiments in primary HSPCs (refer to Figure S1B for a representative example). The final volume of ribonucleoprotein or base editor protein in the buffer with supplement was 17 μL. In experiments with ribonucleoprotein, 1.57μL of chemically-modified sgRNAs (IDT) resuspended at 100 μM in IDTE pH 7.5 (IDT, 11-01-02-02) were pre-complexed for 5-20 minutes with the diluted editor protein. 400,000-500,000 HSPCs were added per electroporation reaction in 5.1 μL of Lonza P3 buffer with supplement. Then, the 17 μL of editor material were added to the cuvette and gently mixed three times. Cells were electroporated using the DZ-100 program in a 4D-Nucleofector X Unit (20 μL cuvettes). Immediately after electroporation, 80 μL of prewarmed media were added to the electroporation cuvette, which was placed in an incubator at 37°C for 5 minutes. Cells were then plated at a density of 500,000 cells/mL in the adequate complete media. In pooled screens, transduced cells were optionally enriched using FACS immediately prior to electroporation (Figure 1A) as described below in the corresponding section. This allowed to reduce the number of electroporations, given the transduction at low MOI of millions of cells.

For the *HBG1/2* promoter screen, we performed two sequential electroporations within 24 hours to increase editing efficiency, with minimal impact on cell viability and differentiation of the cells.

Cas9 protein electroporation

Cells were washed three times in DPBS prior to electroporation, to remove any residual RNAses present in the culture media that could interfere with editing. A master mix of Cas9 ribonucleoprotein was prepared by combining 2.1μL of DPBS, 1.2μL of 100μM sgRNA in IDTE pH 7.5 (IDT) and 1.7μL of 62μM Alt-R S.p. HiFi Cas9 Nuclease V3 (IDT, 1081061), with gentle swirling while pipetting. Following the cell washes described above, cells were resuspended in 20μL of P3 Lonza buffer with supplement (Lonza, V4XP-3032). 5μL of the Cas9 ribonucleoprotein master mix and 1μL of 100μM Electroporation enhancer (IDT) were added to the cells, gently mixed three times, and transferred to an electroporation cuvette. Cells were electroporated using the DZ-100 program in a 4D-Nucleofector X Unit (20 μL cuvettes). Immediately after electroporation, 80 μL of prewarmed media were added to the electroporation cuvette, which was placed in an incubator at 37°C for 5 minutes. Cells were then plated at a density of 500,000 cells/mL in the adequate complete media. In Figure S1D, the condition with Cas9 alone was prepared as described above but the sgRNA was replaced by an equivalent volume of DPBS.

Single-guide RNA design

Single-guide RNAs for screens were designed targeting every possible adenine within positions 4–8 of the protospacer that had a compatible NG PAM using https://github.com/mhegde/base-editor-design-tool. Given that there are many less sgRNAs per variant than in Cas9 KO screens, we decided to keep poly-T and poly-A homopolymers despite the known impairment of synthesis and expression (Figure S1G), and to instead perform filtering during analysis.

For the *HBG1/2* screen, we designed all possible targeting guides with compatible PAM 300 base pairs upstream of the *HBG* promoter. The high homology of this region enables targeting of both promoters using the same sgRNAs. 10 additional sgRNAs from the GecKO v2 library that did not target anywhere in the genome were added.

For the *GATA1* screen, we designed all possible missense-targeting guides covering all exons of ENST00000376670.9, as well as the exon-intron junctions and control non-coding mutations. We also included a number of sgRNAs targeting the 5' UTR of *GATA1*. To produce a more compact library, a fraction of sgRNAs that only mediated silent edits or did not have adenines within the editing window were removed. Thus, our library targets 233 out of the 414 codons of *GATA1* (56.3%). Of these 233 codons, we target missense mutations in 219 and synonymous mutations in 14, several with multiple guides. The number of synonymous mutations editable by ABE8e-Cas9NG is higher (and in fact a fraction of the missense-targeting guides also targets additional synonymous mutations), but we filtered out most of the synonymous-mutation introducing guides from the library to reduce the number of cells required for the screen. 10 additional sgRNAs from the GecKO v2 library that did not target anywhere in the genome were included.

For the *CD33* screens, adenine base editor and cytosine base editor compatible sgRNAs targeting (with an NG PAM) all splice sites were designed using SpliceR v1.14 for transcript ENST00000262262.5.[43] 10 additional sgRNAs from the GecKO v2 library that did not target anywhere in the genome were added, and an sgRNA targeting the chromosome 4 site was included in the ABE screen (Figure S1C).[110] Additionally, we included several sgRNAs targeting the *CD33* gene in non-splice regions as controls.

Lentiviral vector

All screens were performed with an optimized CROP-seq vector. A modified version of the CROP-seq-opti vector[111] was obtained by replacing the 5th base pair following the protospacer with a C instead of a G.[112,113] Additionally, the puromycin resistance cassette was replaced with violet-excited green fluorescent protein (GFP-Vex) to facilitate lentiviral titration and precise FACS-enrichment of infected cells. PspLI (Thermo Fisher Scientific, FERFD0854) and MluI (Thermo Fisher Scientific, FERFD0564) were used to excise the puromycin resistance marker and a gBlock with the GFP-Vex sequence was cloned into the digested vector using the same restriction site overhangs. These modifications simplified lentiviral titration, eliminated potential toxicities associated with antibiotic selection, and allowed for cost-effective and efficient screening of phenotypes even with low numbers of cells. The library assembly protocol was engineered to be compatible with simple and efficient Golden Gate cloning, as described below.

Golden Gate cloning

To assemble lentiviral libraries[114] or individual sgRNA lentiviruses, 58-bp individual oligonucleotides or pools were ordered from Integrated DNA Technologies at a 50 pmol scale (standard desalting), and resuspended at a 10 μM concentration (ordered as GAGCCTCGTCTCCCACCG-[20bp protospacer]-GTTTTGAGACGCATGCTGCA).

An initial extension reaction was performed using the oligo pool (Tables S2–S5) or the individual oligonucleotide, NEB Q5 Hot Start High-Fidelity 2X Master Mix (M0492L) and extension_reaction_primer (Table S1). The following parameters were used for extension: 98°C for 2 minutes; 10 cycles of (64°C for 30 seconds and 72°C for 20 seconds); 72°C for 2 minutes; and hold at 4°C. The product was purified using the QIAquick PCR Purification Kit (QIAGEN, 28104), eluted in 25 μL of water. The vector was digested at 37°C for 4–6 hours with 100 μM DTT and FastDigest Esp3I (Thermo Scientific, FD0454), followed by gel purification with QIAquick Gel Extraction Kit (QIAGEN, 28706X4).

A Golden Gate reaction was prepared using 500 fmol of vector and 10,000 fmol of purified extension reaction product (with the volume required for each calculated using its fragment length and its concentration measured by Nanodrop) with 7μL of FastDigest Esp3I (Thermo Scientific, FD0454), 7μL of T7 Ligase (Qiagen Beverly LLC, L6020L), 20μL of T4 DNA Ligase Buffer (NEB, B0202S) and nuclease-free water to a final reaction volume of 200 μL. The following parameters were used for assembly: 99 cycles of (37°C for 2 minutes and 16°C for 5 minutes), 37°C for 30 minutes and hold at 4°C.

The following morning the product was purified using Zymo DNA Clean & Concentrator- 5 (77001–152) and eluted in 10 μL of water. 2 μL of Purified Golden gate products were transformed into Endura™ Electrocompetent Cells (Biosearch Technologies, 71003-038) using the Biorad Gene Pulser Xcell Total Electroporation System (1652660) with the following parameters: 1.8 kV, 25 μF and 200 Ω. Bacteria were recovered for 20 minutes in the kit's recovery media. 2 μL of bacteria were used to create 4 serial dilutions to evaluate the transformation efficiency and the remaining bacteria were inoculated in 500 mL of LB with 100 μg/ml penicillin/streptomycin (Life Technologies, 15140-122) and grown overnight at 30°C. 16-18 hours later, plasmid DNA was extracted using the NucleoBond Xtra Maxi kit for endotoxin-free plasmid DNA (Macherey-Nagel, 740424.50) and eluted in 200–400 μL of nuclease-free water.

Lentiviral production

293T Human Embryonic Kidney cells (ATCC, CRL-3216) were cultured in 10 cm^2 plates (Corning, 430293) with DMEM (Life Technologies, 1965-118) with 10% FBS (BioTechne, S11550) and 1% penicillin/streptomycin (Life Technologies, 15140-122). Cells were expanded to reach ~80% confluency per plate on the day of transfection. 5–20 10 cm^2 plates were prepared per lentiviral construct.

For each plate, 9 μg of pΔ 8.9 packaging plasmid, 1 μg of VSV.G envelope plasmid, 10 μg of sgRNA vector construct and Opti-MEM media (Gibco, 31985-062) to a final volume of 320 μL were added. The mix was then placed in a Corning® 96-Well Round-Bottom plate (Corning, 38018). 120 μL of FuGENE transfection reagent (Promega, E2691) per plate equivalent were added to the wells while gently swirling the pipette tip and avoiding contact with the walls of the well. The resulting mix was added dropwise to each 293T plate. 12-16 hours later, 293T media was removed and changed to DMEM with 20% FBS and 1% penicillin/streptomycin. 24 hours later the media was harvested in 50 mL tubes (and placed at 4°C), and replaced again with DMEM with 20% FBS and 1% Pen/Strep. 24 hours later the media was harvested again and pooled together with the day 1 harvest.

The viral media was filtered through a Stericup 0.45 μm PVDF membrane (Millipore, SCHVU01RE), and transferred to ultra clear centrifuge tubes (Beckman Coulter, 344058). Virus was subsequently concentrated using a Beckman Coulter SW32Ti Ultracentrifuge with the following parameters: Speed: 24,000 rpm, time: 1 hour and 30 minutes, Temperature: 4°C, maximum acceleration and deceleration 9. The supernatant was removed and the virus pellet was resuspended with the appropriate media. Concentrated virus was stored at -80ºC until further usage.

HSPC lentiviral transduction

Millions of HSPCs were transduced at a density of 500,000-1 million cells per mL 1–2 days after thawing (Figure 1A). Concentrated virus was added to CD34+ HSPCs along with 8 μg/mL polybrene (Sigma Aldrich, TR-1003-G). Cells were then spinfected at 2,000 rpm for 90 mins at 37ºC. 12-16 hours after spinfection, the media was replaced by the appropriate complete media. We targeted low MOIs (10-25% of transduced cells) and relied on FACS enrichment of transduced cells prior to single-cell genomics applications (Figures 1A and S3G).

We always ensured we had excess sgRNA coverage to avoid dropout in the screens. We used high coverage for functional screens (>3000x for CD33, >7,500x for HBG1/2 and >3,000x for GATA1) at every step. This number can be likely reduced to approximately 1,000 cells per sgRNA while still maintaining adequate coverage for the majority of the guides. A critical procedure that we implemented was the sorting of infected cells prior to electroporation. This procedure is not absolutely necessary, but it allows for the transduced cells to be enriched by over 5-fold.

Flow cytometry sample preparation

For extracellular stains, cells were spun at 400 x *g* for 5 minutes and washed twice with FACS buffer (PBS with 1% FBS). Cells were stained with conjugated antibodies in 100 μL of FACS buffer for 30 minutes in the dark at 4ºC (key resources table). Following incubation, cells were washed twice in FACS buffer and resuspended at an appropriate density for analysis or sorting.

For FACS enrichment prior to electroporation (Figure 1A), cells were simply spun down at 370 x g for 5 minutes and resuspended in PBS. For FACS enrichment prior to single-cell genomics, cells were resuspended in PBS + BSA 0.05% (Sigma-Aldrich, A9418).

For the intracellular fetal hemoglobin stain for the functional screen and the validation experiments (Figure 4), cells were washed with PBS-0.1% BSA (Sigma-Aldrich, A9418) and fixed in 500 μL of cold 0.05% glutaraldehyde (Electron Microscopy Sciences, 16121) for 10 minutes at room temperature, with vortexing after glutaraldehyde addition. Cells were then washed 3 times with 2 mL of PBS-0.1% BSA. The pellet was resuspended by vortexing in 0.5 mL 0.1% Triton X-100 (Sigma-Aldrich, X100–1L) and incubated for 5 minutes at room temperature. Cells were washed once with 2 mL of PBS-0.1% BSA, resuspended in 70 μL with 2 μL of Fetal Hemoglobin Monoclonal Antibody APC (Life Technologies, MHFH05) per 500,000 cells and incubated for 15 minutes at room temperature. Cells were washed twice with 2 mL of PBS-0.1% BSA and resuspended in the appropriate volume of FACS buffer prior to sorting.

Fluorescent-activated cell sorting (FACS)

To maximize recovery of cells, 1.5mL or 5 mL low-bind tubes (Eppendorf, 0030122356) were coated with 1mL of sterile FACS buffer (PBS +1% FBS) or PBS+0.05% BSA for single-cell genomic experiments and set aside. Cells were sorted using a Sony MA900 instrument. Following sorting, tubes were spun at 450 x *g* for 5 minutes. Cell pellets were either resuspended in media for continued culture, flash frozen for prolonged storage at –80ºC or processed for genomic DNA extraction.

Editing efficiency

Genomic DNA (gDNA) was extracted from edited cells using Qiagen Micro or Mini kits, depending on the number of input cells (QIAamp DNA Micro Kit: 56304 and QIAamp DNA Mini Kit: 51304). For very low input cell numbers (e.g. certain cell sorts of rare populations), carrier RNA was included during gDNA extraction and elution volumes were reduced to the minimum recommended. NCBI's PrimerBLAST was used to design primers specific to the target locus for an amplicon of 150–350bp of size, with the forward primer less than 100bp away from the targeted location (Table S1).[115] An initial pre-amplification PCR was performed on 100ng of gDNA, which were added to a mixture of NEB Q5 Hot Start High-Fidelity 2X Master Mix and primers following the New England Biolabs protocol for a 25 μL PCR reaction. The following parameters were used for PCR: 98ºC for 30 seconds; 24–28 cycles of (98ºC for 20 seconds, annealing temperature for the primer pair for 20 seconds, 72ºC for 20 seconds); 72ºC for 2 minutes; and hold at 4ºC. The optimal annealing temperature was determined using New England Biolabs Tm calculator. The optimal number of cycles was determined using quantitative PCR (qPCR). The qPCR mixture contained 12.5μL Q5 2X Master Mix, 1.25μL of 10μM forward primer, 1.25μL of 10μM reverse primer, 2.5μL of SYBR green 10X (VWR International LLC, 12001-796), 100ng of gDNA and water up to 7.5μL.

The 25μL PCR product was diluted to 50μL, from which 40μL were purified using one round of 1X SPRI beads (Beckman Coulter, B23318) and eluted in 50μL of water. Illumina universal adapters were added in a final PCR reaction. 3μL of the purified PCR produced were added to a mixture of 5μL Q5 2X Master Mix, 1μL of 5μM uniquely indexed P5 universal Illumina adaptor and 1μL of 5μM uniquely indexed P7 universal Illumina adaptor, using the following parameters: 98ºC for 30 seconds; 4-10 cycles of (98ºC for 10 seconds, annealing temperature for the primer pair for 20 seconds, 72ºC for 20 seconds); 72ºC for 2 minutes; and hold at 4ºC. The optimal number of cycles was again determined using qPCR. PCR products were purified using one round of 1X SPRI beads. The concentration and size of the libraries were assessed using the Agilent 2100 Bioanalyzer High Sensitivity DNA kit (Agilent

Technologies, 5067-4626 and 5067-4627). If heteroduplex PCR-bubble products were identified on the Bioanalyzer, the number of cycles of the final PCR reaction was reduced. Libraries were pooled and sequenced in an Illumina Miseq system (with Nano, Micro, or regular flow cells depending on the number of samples) generally using a 150-160 base pair (Read 1), 150-160 base pair (Read 2), and 8 base pair (Index 1) configuration (with variations for certain amplicons).

Functional screen sample processing

Genomic DNA was extracted and processed as described in the previous section, with the following modifications. For functional screens, all available genomic DNA was processed in 96-well plates, using 100ng of DNA per PCR reaction. Upon completion the reactions for each sample were then pooled and purified using SPRI beads. In the *HBG1/2* functional screen, multiple PCR2 were performed per PCR1 pool, which were treated as technical replicates for analysis. To increase library diversity during sequencing, we used a staggered forward primer targeting the U6 promoter and a reverse primer hybridizing to the vector's scaffold (Table S1). This same strategy was employed to assess library diversity from cloned library plasmids prior to lentiviral production. Libraries were sequenced in an Illumina Miseq, Nextseq 500 or Nextseq 2000 instrument depending on the number of samples and desired read-depth.

For cells fixed for intracellular fetal hemoglobin analysis, the Quick-DNA FFPE kit (Zymo, D3067) was used to extract gDNA directly after sorting. Cells were not subjected to the deparaffinization solution step and samples were digested overnight at 55ºC for 12-16 hours. Beta-mercaptoethanol (Sigma-Aldrich, M6250) was included in the genomic lysis buffer at a concentration of 0.5%. To maximize DNA yield, the elution water was heated to 60ºC and two 50 μL elutions were performed per sample.

Real-time quantitative PCR (RT-qPCR)

Total RNA was extracted using the AllPrep DNA/RNA Mini Kit (QIAGEN, 80204) or the RNeasy Mini Kit (QIAGEN, 74106). RNA was reverse transcribed using the iSCRIPT cDNA Synthesis kit (Bio-Rad, 1708891) following manufacturer's instructions. RT-qPCR was performed using iQ SYBR Green Supermix (Bio-rad, 1708880) and the CFX384 Touch Real-Time PCR Detection System (Bio-Rad). Primers used are listed in Table S1.

Full-length RNA sequencing for isoform analysis

Total RNA extracted from edited erythroblasts on day 9 post-electroporation was processed using the SMART-Seq v4 ultra low Input RNA Kit for Sequencing (TAKARA, 634889) and sequenced on a Nextseq 2000 instrument.

Morphological analysis of primary cell cultures

To analyze the morphology of base-edited differentiating primary cell cultures, 100,000 cells were harvested, washed and centrifuged using a Cytospin 4 centrifuge (Thermo Scientific) at 500rpm for 5 minutes with low acceleration. Air-dried slides were stained using May-Grünwald solution (Sigma Aldrich, MG1L) for 5 minutes, rinsed 4 times for 30 seconds in water, and stained using Giemsa solution (Sigma Aldrich, 32884) for 15 minutes. Slides were examined using a Mica instrument (Leica Microsystems).

Transplantation of HSPCs into NBSGW mice

For xenotransplantation assays NBSGW mice were obtained from Jackson Laboratory (026622).[48] Littermates of the same sex (4 males and 5 females) were randomly assigned to experimental groups, and were 8 weeks old at the time of transplantation. All animal experiments were approved by the Boston Children's Hospital Institutional Animal Care and Use Committee (A3303-01).

Mice of 8 weeks of age were injected with healthy newborn umbilical cord blood-derived CD34+ HSPCs. Discarded cord blood units were obtained from the Pasquarello Tissue Bank at Dana-Farber Cancer Institute (IBC-P00000180). Umbilical cord blood-derived CD34+ cells were isolated using the EasySep™ Human Cord Blood CD34 Positive Selection Kit II (StemCell Technologies, 17896) cultured in StemSpan SFEM II human hematopoietic stem cell expansion media (StemCell Technologies, 02690) with 1% L-glutamine (Thermo Fisher, 25-030-081) and 1% Penicillin/Streptomycin (Life Technologies, 15140-122) and supplemented with the CC100 cytokine cocktail 100x (StemCell Technologies, 02690), 50 ng/mL TPO (PeproTech, 300-18) and 35 μM UM171 (StemCell Technologies, 72912).

200,000-300,000 CD34+ cells per mouse were injected via tail vein 2-3 days following base editor ribonucleoprotein electroporation. To monitor engraftment, peripheral blood was sampled at 6-, 12- and 16-weeks post-transplantation by retro-orbital sampling. At 16 weeks post transplantation, animals were sacrificed, and bone marrows and spleens were collected. Bone marrow cells were collected by flushing of bilateral femurs and tibias. Human chimerism was assessed by flow cytometry using anti-human-CD45 and anti-mouse-CD45 antibodies.

Cell type composition and CD33 protein levels were additionally measured using 10X Genomics scRNA-seq with Feature Barcoding technology. A TotalSeq™-B0052 (BioLegend, 366635) anti-CD33 antibody was added to the cells. Cells were then processed using v3.1 Chemistry Dual Index kits with Feature Barcoding technology (10X Genomics). Sequencing reads were aligned to a combined mouse and human reference genome using Cellranger *count*, and mouse cells were filtered out. Gene expression matrices were normalized using Seurat v4 *NormalizeData*, which performs log transformation of counts scaled by a factor of 10,000. Standard processing of the data with *FindVariableFeatures, ScaleData, RunPCA, FindNeighbors* (using 40 dimensions), *FindClusters* and *RunUMAP* was performed.

Perturb(BE)-seq

Following final FACS enrichment to remove debris and purify transduced cells, 5-mL PBS-BSA coated tubes described above were spun down at 400 x g for 5 minutes at 4ºC. The entire supernatant was carefully removed and 1mL of PBS+0.05% BSA was added. Cells were then counted twice using an automated Countess 2 cell counter (Thermo Fisher, I-CACC2). Cell viability as determined by Trypan Blue was always >95%. Cells were again washed with 5mL PBS+0.05% BSA and resuspended at a final concentration of 1,000 cells / μL and kept on ice. Single-cells were immediately processed using v3.1 Chemistry Dual Index kits according to manufacturer's instructions (10X Genomics), using 20μL of cell suspension and 23.2 μL of water on the cell suspension loading step. Libraries were sequenced using a 28 base pair (Read 1), 10 base pair (Index 1), 10 base pair (Index 2), 90 base pair (Read 2) configuration on Nextseq 2000 or Novaseq S4 instruments.

For the arrayed CD33 Perturb(BE)-seq experiment (Figure 3), each of the four conditions were transduced and electroporated separately. Following final FACS enrichment, a similar number of cells from each of the conditions were pooled together. A TotalSeq™-B0052 (BioLegend, 366635) anti-CD33 antibody was added to the cells, using the 10X Genomics Cell Surface Protein Labeling "Wash Protocol 1". Cells were then processed using v3.1 Chemistry Dual Index kits with Feature Barcoding technology (10X Genomics).

Notably, we were able to detect enough CROP-seq reads to assign single-cell perturbations in 72.8% of cells even without PCR enrichment at only ~20,000 reads per cell in gene expression libraries (Figures S3C and S3D). PCR enrichment increased the CROP-seq transcript counts by 15-fold per cell and allowed CROP-seq transcript detection in 94.8% of cells. Importantly, the dominant perturbation was still concordant in 95.6% of the cells in which CROP-seq transcripts were detected with both strategies (Figures S3D–S3F). In terms of the implications of the workflow, it only adds approximately one day to the protocol, so in practice we recommend always performing the PCR enrichment, which might be especially beneficial in screens performed at high MOIs. To enrich CROP-seq polyadenylated transcripts containing the identity of the perturbation, or other transcripts of interest (e.g. *GATA1*) we adapted the PCR-based enrichment strategy reported in Al'Khafaji et al.[116] Following the initial enrichment PCR1 with primer F_CROPseq_PCR1 (Table S1) and AAO272, we performed PCR2 reactions using 15μL of Q5 Hot Start High-Fidelity 2X Master Mix (NEB, M0492L), 1.25 μL 25μM of each uniquely indexed universal Illumina adaptor and 12.5μL of PCR1 product, using the following parameters: 98ºC for 30 seconds; 28 cycles of (98ºC for 15 seconds, 69ºC for 15 seconds, 72ºC for 20 seconds); 72ºC for 2 minutes; and hold at 4ºC. PCRs for the same sample were then pooled, purified using one round of 1X SPRI beads (Beckman Coulter, B23318) and eluted in EB buffer, which was quantified using the Agilent 2100 Bioanalyzer High Sensitivity DNA kit (Agilent Technologies, 5067-4626 and 5067-4627). Libraries were sequenced using a 28 base pair (Read 1), 8 base pair (Index 1), and 20-42 base pairs (Read 2) configuration on Nextseq 500 or Nextseq 2000.

Massively parallel single-cell pooled genotyping

To couple the identity of each sgRNA with the base edits it mediates at endogenous loci in primary HSPCs, 30 non-overlapping amplicons (ranging 189–265bp sizes) targeting genomic DNA GATA1 exons and exon-intron junctions, *AAVS1, CD33*, and the *HBG1/2* promoters were designed (Table S1). We also included a primer pair targeting the sgRNA sequence and 6 additional primers targeting other regions of the integrated CROP-seq lentiviral vector. This way, for each single-cell, we would profile amplicons containing the sgRNA information as well as the mutational status at the target locus.

As a proof-of-concept for this strategy, we transduced the HBG1/2 and GATA1 ABE systematic mutagenesis libraries and electroporated ABE into HSPCs. For GATA1, HSPCs were maintained in HSPC stem cell expansion media supplemented with 35μM UM171, as described above, to reduce dropout of deleterious mutations impairing erythropoiesis. Following final FACS enrichment to remove debris and purify transduced cells, 5-mL PBS-BSA coated tubes described above were spun down at 400 x g for 5 minutes at 4ºC. The entire supernatant was carefully removed and 1mL of PBS was added. Cells were then counted twice using an automated Countess 2 cell counter. Cell viability as determined by Trypan Blue was always >95%. Cells were again washed with 5mL PBS and resuspended at a final concentration of 4,000 cells / μL in Cell Buffer kept on ice. Cells were immediately processed using Tapestri Single-Cell DNA Sequencing V2 (Mission Bio) following manufacturer's instructions. Libraries were sequenced using a 150 base pair (Read 1), 8 base pair (Index 1), 8 base pair (Index 2), 150 base pair (Read 2) configuration on Nextseq 500 or Nextseq 2000 instruments.

Two patients with GATA1 variants of unknown significance

The first patient (XY male) carrying the c.220+2T>C (chrX:48791331T>C) *GATA1* VUS was first found to be anemic at 10 months during a pre-operative evaluation. At 3 years of age, remarkable findings included anemia with hemoglobin 6.8 g/dL (normal range 11–13.7 g/dL) and MCV 79 fL (normal range 75–86 fL) and thrombocytosis with 1,021 x 10^9 platelets/L (normal range 150–450 x 10^9 platelets/L). A bone marrow biopsy revealed a reportedly normocellular marrow with dyserythropoiesis. No cytogenetic abnormalities were noted. A hemizygous T>C variant at the c.220+2 position (chrX:48791331) of *GATA1* was identified, and the patient received the diagnosis of congenital dyserythropoietic anemia. The patient remained transfusion-dependent every 3–5 weeks with resultant transfusion-related iron overload. On follow-up at 16 years of age, persistent anemia with hemoglobin 7.9 g/dL (normal range 12.4–16.4 g/dL) and MCV 84 fL (normal range 80–96 fL), thrombocytopenia with 115 x 10^9 platelets/L (normal range 150–450 x 10^9 platelets/L), and elevated erythropoietin levels of 2,823 mU/mL (normal range 2.6–18.5 mU/mL) were noted. A bone marrow aspirate at 16 years of age revealed erythroid hypoplasia and dysplastic features in megakaryocytes. Cytogenetics revealed a new appearance of

monosomy 7 in 15% of interphase nuclei. The patient was enrolled in the clinical trial NCT02720679 and is undergoing further workup for potential stem cell transplantation.

The second patient (XY male) carrying the c.218C>T (chrX:48791327C>T) *GATA1* VUS was referred with anemia at 12 months old. Hemoglobin ranged 7.5–8.3g/dL (normal range 11–13.5 g/dL) and MCV was 90.1-96.8 fL (normal range 73–85 fL) with no known family history of the condition. Adenosine deaminase levels of 1,350 mU/g Hb (normal $\leq$ 1,000mU/g Hb) and erythropoietin levels of 1,245 mU/mL (normal range 2.6–18.5 mU/mL) were detected. Bone marrow biopsy revealed hypercellularity with atypical megakaryocytic hyperplasia, erythroid hypoplasia, and some dyserythropoietic features (Figure S7A). A normal karyotype was noted. At 22 months of age the bone marrow biopsy was repeated and remained unchanged. An extended sequencing panel including commonly mutated genes in Diamond-Blackfan anemia revealed the c.218C>T variant in *GATA1*. Steroid therapy was initiated after a descent in height percentile was noted concomitant with worsening anemia (hemoglobin of 6.8 g/dL). This treatment resulted in normalization of hemoglobin levels for his age (11.2 g/dL) and thrombocytosis (997 x 10^9 platelets/L). The patient remains on corticosteroid therapy.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests are indicated in the figure legends. Statistical significance is denoted in Figures using the following notation: ns, not statistically significant; ****, $p < 0.001$; ***, $p < 0.001$; **,$p < 0.01$; *,$p < 0.05$. All error bars represent standard deviation.

Analysis of editing efficiencies

Raw fastq.gz files were demultiplexed using the bcl2fastq v2.20 conversion software. Editing efficiency analysis was performed using CRISPResso2.[117] In Figure S1D, editing efficiencies were corrected by the percentage of cells transduced with the vector, measured by flow cytometry, as transduced cells in this experiment were not FACS-purified prior to gDNA extraction.

Analysis of plasmid library diversity

Demultiplexed fastq.gz files were loaded using the readFastq package and the sgRNA sequencing flanking "CACCG" and "GTTT" was extracted. The small percentage of sgRNAs that were present at low levels were filtered out.

Analysis of functional base editing screens

For the *CD33* and *HBG1/2* screens, demultiplexed fastq.gz files were quantified using poolQ v3.3.2. Technical replicates were first quantified individually and then their reads were combined after confirming consistent behavior across them. The log normalized counts for each file were computed as $\log_2$(number of reads for a given sgRNA / number of reads of all sgRNAs in the sample $\cdot$ 10^6 + 1). We decided to keep poly-T and poly-A homopolymer sgRNAs on the original designs despite the known impairment of synthesis given that some could still perform well on the screen (Figure S1G), and instead performed filtering at this stage. sgRNAs with low frequencies across conditions were also filtered out. To compute enrichment and depletion among groups, the difference of the log normalized counts of the two conditions was computed. Then, the resulting $\log_2$(fold change) was z-scored as follows:

$$\frac{log_2(fold\ change\ individual\ sgRNA) - mean(log_2(fold\ change\ all\ sgRNAs))}{sd(log_2(fold\ change\ all\ sgRNAs))}$$

Pre-processing of Perturb(BE)-seq data

Raw bcl files were demultiplexed using Cellranger v6.0.1 *mkfastq* and gene expression matrices were obtained using Cellranger *count*. CROP-seq transcript reads were either aligned to a custom reference with the sequence of the sgRNAs, or processed using Cellranger *count* as "crispr" feature barcoding samples, with similar results. The custom reference was created using STAR with the –genomeSAindexNbases 7 parameter, using 126 bases upstream of the scaffold, the scaffold sequences, and 60 base pairs downstream of the scaffold. GATK *CreateSequenceDictionary* was run on the resulting fasta custom genome file and STAR alignment was performed using default alignment parameters. Read names and the identity of aligned sgRNAs were extracted from the resulting bam files and used to subset R2 and R1 fastq.gz files using Seqtk *subseq*. From each R1, the first 16bp were extracted to recover the cell barcode (CB) identity, and the following 12bp were assigned to the UMI. The resulting matrices (with CB, UMI, and sgRNA) were deduplicated using the UMI and filtered using the CB present in the gene expression matrices. When processed using Cellranger *count*, CB and UMI error correction were additionally performed. CD33 barcoded-antibody counts were also quantified using Cellranger *count*.

The dominant sgRNA for each single-cell was assigned to the sgRNA with the highest number of CROP-seq transcript counts in a cell. We assign a given sgRNA to a cell if it has >1.3x more counts than the second sgRNA with most counts in that cell. As shown in Figure S3F, this strategy agrees well on the assignment of the relevant sgRNA in a cell both using reads from the enrichment-PCR libraries and scRNA-seq reads.

Gene expression matrices were normalized using Seurat v4 *NormalizeData*, which performs log transformation of counts scaled by a factor of 10,000. Standard processing of the data with *FindVariableFeatures, ScaleData, RunPCA, FindNeighbors* (using 40 dimensions), *FindClusters* and *RunUMAP* was performed.

Perturb(BE)-seq analysis (arrayed experiment)

We detected cells with the lowest levels of CD33 protein expression in each cluster by scaling the CD33 protein counts in each cell by the average CD33 levels of the cluster they belonged to (given that, for instance, erythroid cells express low baseline levels of CD33).

Perturb(BE)-seq analysis (HBG1/2 screen)

We defined the scaled (*HBG1+HBG2*) levels for each single-cell using the natural logarithm of 1+x (*log1p* function) of the sum of *HBG1* and *HBG2* counts divided by the total number of unique molecular identifiers sequenced in each single-cell and scaled by a factor of 10,000, which we plotted as cumulative distributions in Figure S4B. We fitted a linear model using the lm package in R in which the sgRNA identity in single-cells is used as a predictor for the observed number of single-cell scaled counts for *HBG1+HBG2* in the Perturb(BE)-seq screen:

$$(scaled\ HBG1\ +\ HBG2\ counts)\ \sim\ sgRNA$$

Analysis of sgRNA depletion (GATA1 screen)

We compared the distribution of each sgRNA in the day 9 erythroid differentiation timepoint (total of 13 days in culture) with respect to transduced, non-electroporated cells to identify variants critical for GATA1 function. We filtered out sgRNAs with low representation in transduced, non electroporated cells. In Figures 5C and S6B we queried the Genome Aggregation Database (gnomAD v2.1.1) for reported GATA1 mutations across 125,748 exome and 15,708 whole-genome sequences. We plotted missense and partial loss-of-function mutations, totalling in 158 variants. Corresponding allele counts were graphed based on variant position.

Evolutionary conservation across the *GATA1* locus was assessed using conservation scoring by PhyloP100way (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phyloP100way/) from the PHAST package (http://compgen.bscb.cornell.edu/phast/). In short, PhyloP100way contains conservation measures for individual nucleotides based on multiple sequence alignments of 100 vertebrate species. Larger PhyloP values denote highly conserved nucleotides across species.

Analysis of sgRNA lineage enrichment

An erythroid score was defined to quantify enrichment of sgRNAs across blood lineages. In brief, the erythroid score for a particular sgRNA was defined as the *Z*-scored ratio of the number of cells in erythroid clusters to the cells in all other clusters (Figure S4D). sgRNAs were then ranked by erythroid score and annotated with mutation consequence information using Variant Effect Predictor (useast.ensembl.org/info/docs/tools/vep). Pymol was used to visualize sgRNAs targeting the c-terminal zinc finger region. The crystal structure of murine GATA1 zinc-fingers complexed to DNA (PDB ID: 3VD6) was used as a template, which only differs in one amino acid with respect to that region of human GATA1.

Pooled single-cell genotyping analysis

Sequenced fastq.gz files were demultiplexed using Illumina's BCL Convert v4.0.3 CLI and processed using the Tapestri Pipeline V2, which includes adapter trimming, sequence alignment with BWA, cell barcode correction, cell identification and variant calling with HaplotypeCaller from GATK v4. Some fastq.gz files were downsampled prior to input into the pipeline to avoid excessive read coverage, which results in diminishing returns during mutation calling.

To identify the dominant sgRNA in each cell, BWA alignments to the lentiviral sgRNA sequences were obtained, and protospacer sequences were extracted. Using the read name associated with those alignments, each cell barcode sequence was paired with the corresponding lentiviral sgRNA sequence, analogously to the strategy performed above in Perturb(BE)-seq data.

The loomR package (https://github.com/mojaveazure/loomR) was used to extract cell barcodes passing filters processed with the Tapestri Pipeline V2. We then focused on cells that had called variants concordant with the dominant sgRNA present in the cell.

Full-length bulk RNA sequencing analysis

Sequenced fastq.gz files were demultiplexed using Illumina's BCL Convert v4.0.3 CLI and quantified using Salmon v1.10.0 to the hg38 reference transcriptome.[118] Quality control was performed using principal component analysis of the samples, followed by calculation of the transcript per million counts as well as the ratio of full-length and short *GATA1* isoforms.

Transcriptional signature of *GATA1* mutants

Differential gene expression using a Wilcoxon Rank Sum test was performed between cells with the lowest functional scores from the screen and control sgRNAs, to identify a *GATA1* perturbational gene expression signature that was then used to cluster all sgRNAs. The input for clustering was the average expression (using scaled counts) for each gene belonging to the aforementioned signature across single cells sharing the same sgRNA. Clustering was performed using Euclidean distance, and plotted using the *pheatmap* package in R.

RNA velocity and vector field reconstruction

Spliced and unspliced counts were obtained from demultiplexed fastq.gz files using kb-python (https://github.com/pachterlab/kb_python). Velocity analysis was then performed using *Dynamo*.[78] To improve downstream analyses, we appended genes

previously used in Dynamo for hematopoiesis datasets and differentially expressed genes between *GATA1* perturbations and controls. These genes were then combined with highly variable genes identified by Dynamo to perform PCA (Principal Component Analysis), followed by UMAP (Uniform Manifold Approximation and Projection for Dimension Reduction), using default settings. To account for the Multiple Rate Kinetics genes and correct RNA velocity flow, we used the dynamo.tl.gene_wise_confidence function to filter genes whose expression kinetics did not follow clockwise dynamics on the spliced-unspliced RNA phase plane for the HSPC to Erythroid progenitor transition, as well as the Erythroid progenitor to Erythroid transition. To visualize RNA velocity flow, we projected the corrected RNA velocity to UMAP space and used dyn.pl.streamline_plot to generate the streamline plot. The gaussian kernel density estimates were plotted using Scanpy on filtered UMAPs with the strongest GATA1 perturbations and non-targeting control sgRNAs.[119]

CellPress
OPEN ACCESS

Supplemental figures

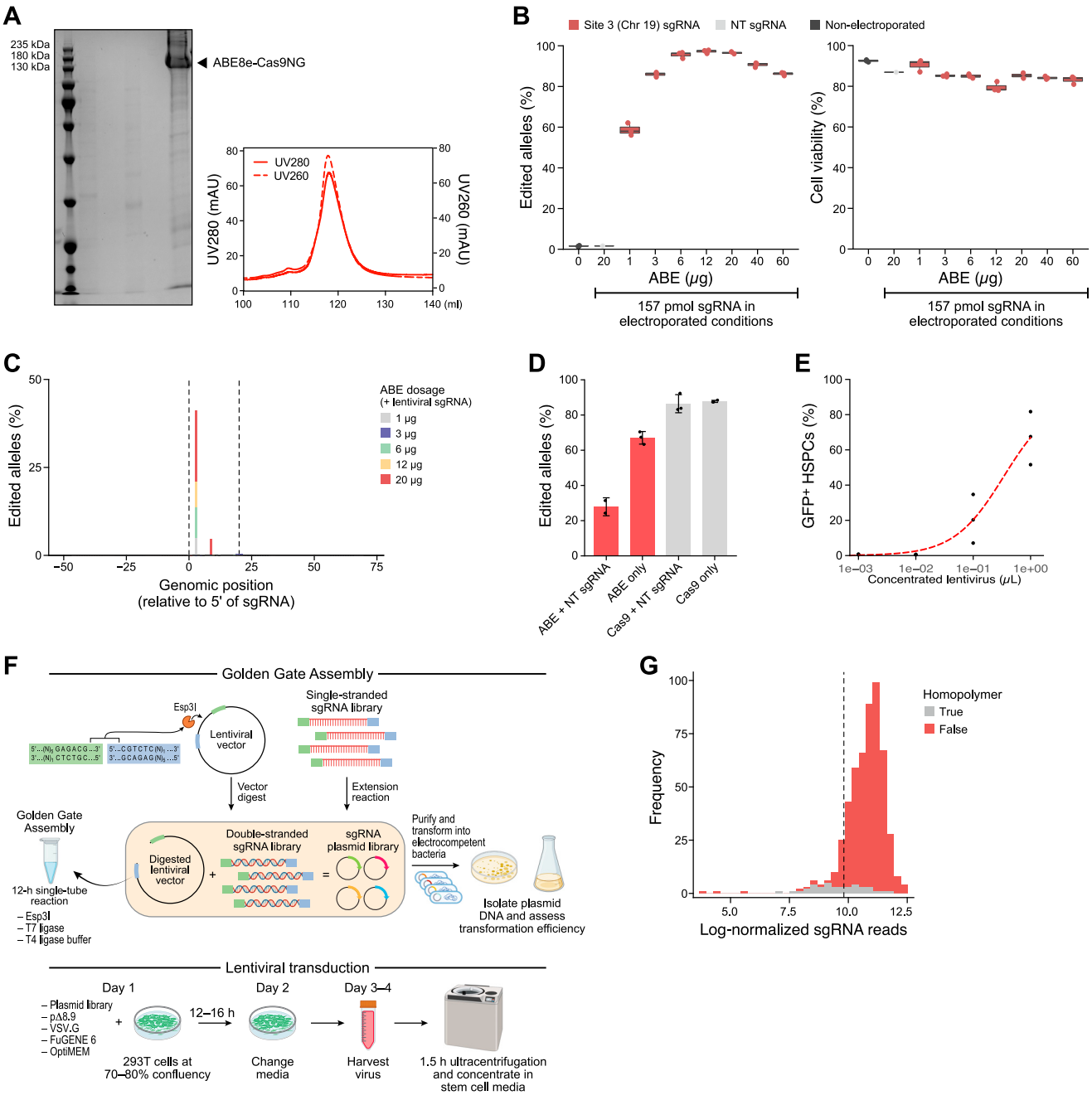


Figure S1. Base editor protein purification and editing, and lentiviral library production and transduction, related to Figure 1

(A) Left, representative polyacrylamide gel electrophoresis of purified adenine base editor (ABE8e-Cas9NG). Right, quantification of the UV280 and UV260 ratios for ABE8e-Cas9NG.

(B) Left, percentage of edited alleles as a function of ABE8e-Cas9NG concentration precomplexed with fixed concentration of an sgRNA targeting site 3 (chromosome 19) and a non-targeting sgRNA, or not electroporated. n = 3 independent electroporations. Right, percent of viable cells for each condition.

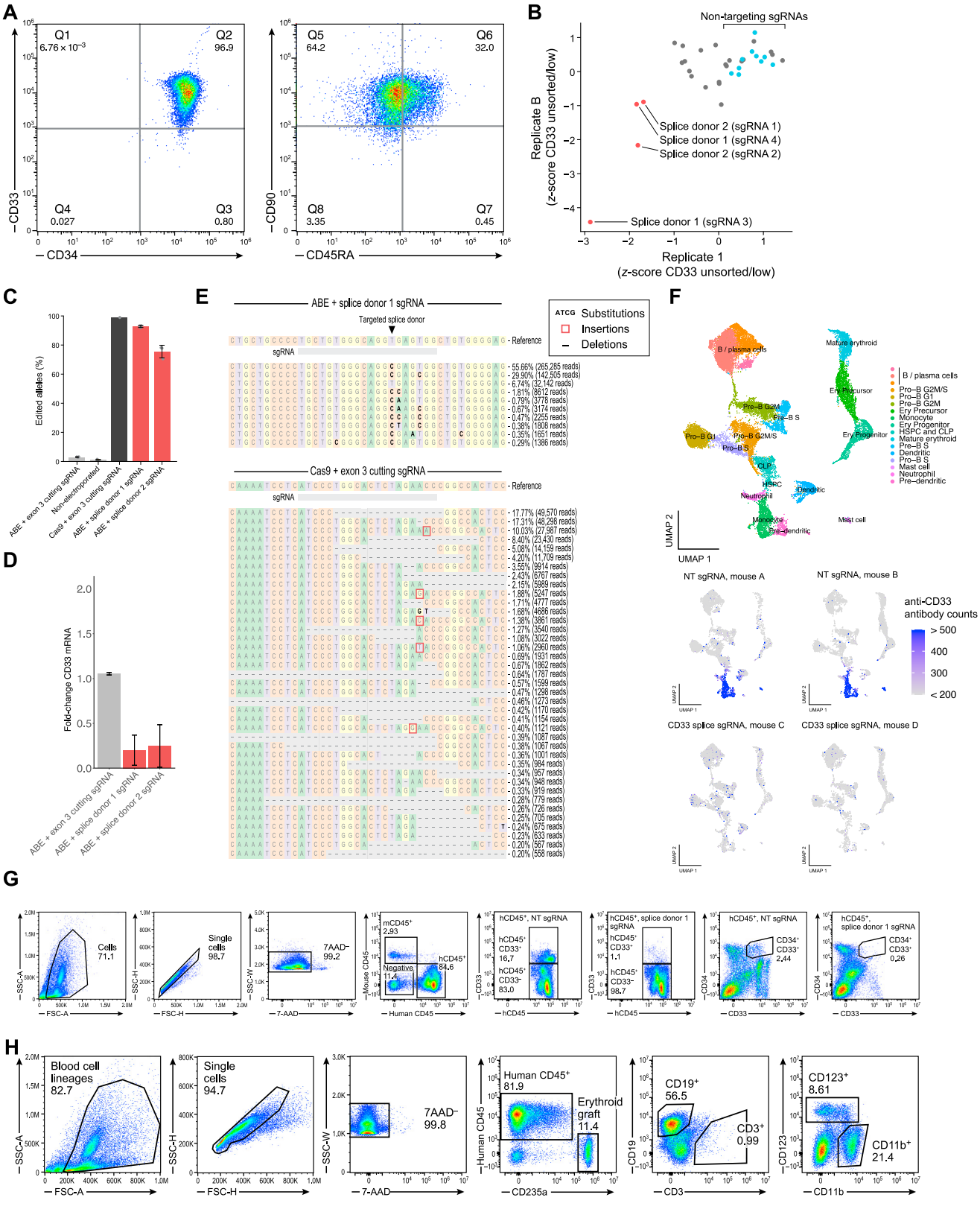
(C) Percentage of edited alleles using lentivirally transduced sgRNAs on site 5 (chromosome 4), as a function of ABE protein dosage in primary HSPCs.

(D) Percentage of edited alleles of lentivirally transduced sgRNAs with and without ABE8e protein (at a dosage of 20 μg) precomplexing with non-targeting sgRNAs prior to electroporation. The percentage of edited alleles with Cas9 targeting site 6 (chromosome 19) as a function of pre-complexing is shown for comparison. n = 2–3 independent electroporations.

(legend continued on next page)

(E) Percentage of GFP+ infected hematopoietic stem and progenitor cells as a function of lentivirus concentration. The curve shows a Michaelis Menten model fitted to the data. n = 3 independently produced viruses and infections.
(F) Top, schematic of the Golden Gate assembly strategy employed for rapid and high diversity sgRNA library assembly. Bottom, schematic of the strategy to obtain high-titer lentivirus. Some schematics were created with BioRender.
(G) Distribution of sgRNA $\log_2$(sequencing reads per sgRNA/total sgRNA reads $\times 10^6 + 1$) from a 524-guide pooled library. Poly(T) and poly(A) homopolymer-containing sgRNAs are highlighted. The dashed line denotes −1 standard deviation.

CellPress
OPEN ACCESS



(legend on next page)

Figure S2. Identification and characterization of a CD33 splice site single-base edit that results in CD33 ablation, related to Figure 2

(A) CD34, CD33, CD90, and CD45RA FACS plots of hematopoietic stem and progenitor cells (HSPCs) on day 3 post-thawing.

(B) *Z* scored $\log_2$(FC) in single-guide RNA (sgRNA) reads between HSPCs with the bottom 10% CD33 levels and the unsorted population, shown for two independent donor and plasmid libraries. Non-targeting sgRNAs are highlighted in blue and outlier splice-site-targeting sgRNA are shown in red.

(C) Editing efficiency of each of the conditions shown in Figure 2C.

(D) Validation of the two top hits of the screen by quantitative real-time PCR of *CD33* mRNA. Error bars represent the standard deviation for 2–3 independent electroporations, with 3 technical replicates each.

(E) Top, percentage of each of the alleles detected in HSPCs edited with ABE8e and the CD33 splice donor 1 sgRNA. Bottom, percentage of each of the alleles detected in HSPCs edited with Cas9 and an exon 3 sgRNA.

(F) Top, combined UMAP of the single-cell RNA sequencing profiles of human cell types recovered from the bone marrow of four mice 16 weeks post-transplantation. Two mice were transplanted with cells electroporated with ABE8e and non-targeting sgRNA, and two mice were transplanted with ABE8e and splice donor 1 sgRNA. Bottom, separate UMAPs for each mouse. The color scale denotes the levels of CD33 protein of each single cell, measured with barcoded antibodies.

(G) Representative flow-cytometry density plots of the gating strategy used to analyze CD33 and CD34 levels in the transplant of ABE-edited human HSPCs into NBSGW mice. For the last gate, representative plots from cells electroporated with ABE8e and non-targeting sgRNA or ABE8e and splice donor 1 sgRNA are shown.

(H) Representative flow-cytometry density plots of the gating strategy used to analyze CD235a (GYPA), CD123, CD11b, CD19, and CD3 levels in the transplant of ABE-edited human HSPCs into NBSGW mice.

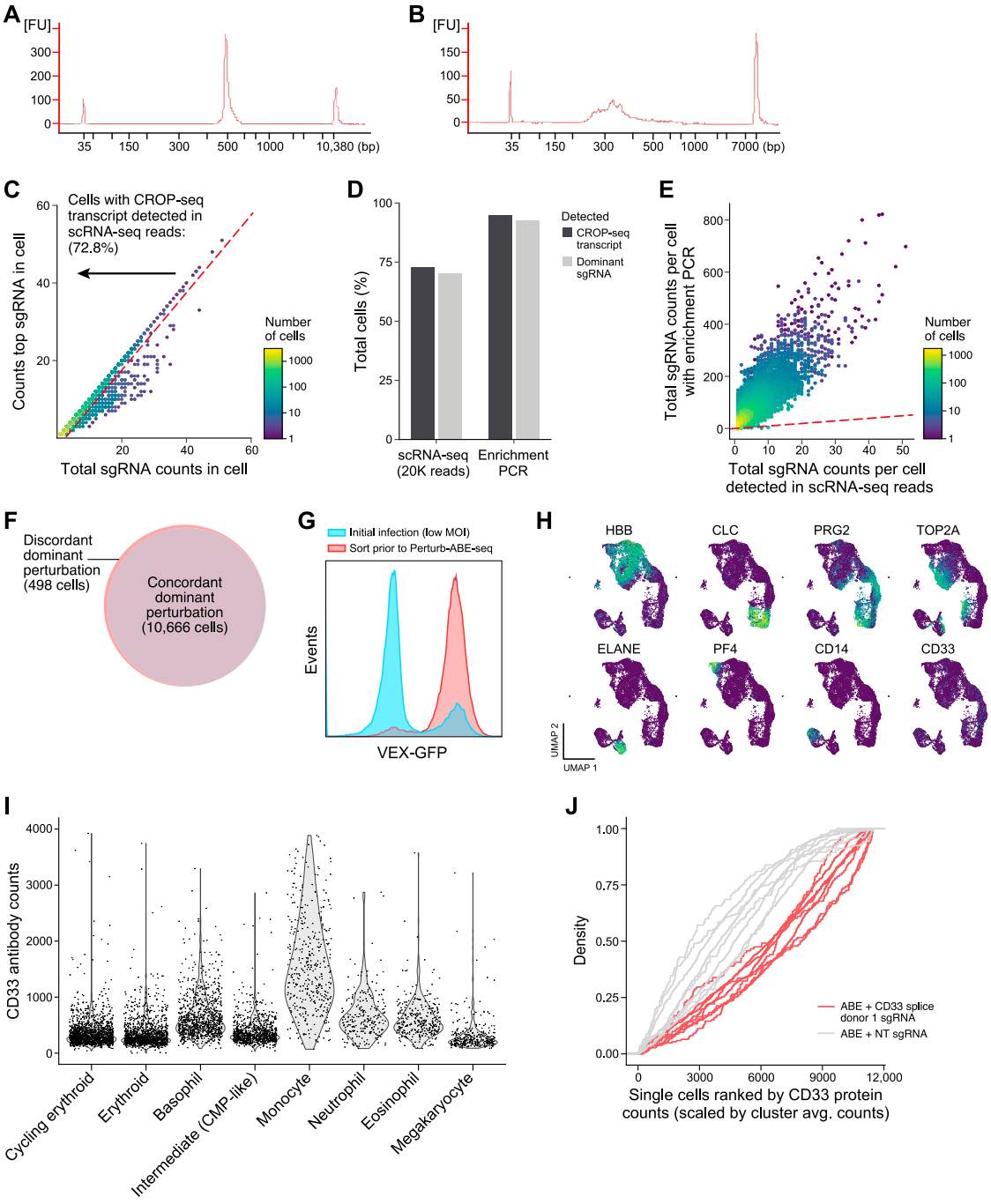


Figure S3. Characterization of single-cell gene screens in primary human hematopoiesis, related to Figure 3

(A) Bioanalyzer trace of an enrichment PCR library performed on scRNA-seq cDNA from cells transduced with the modified CROP-seq vector.

(B) Bioanalyzer trace of a direct sgRNA capture library from HSPCs transduced with a vector containing capture sequence 2 in the 3′ end of the sgRNA, lacking a single defined peak.

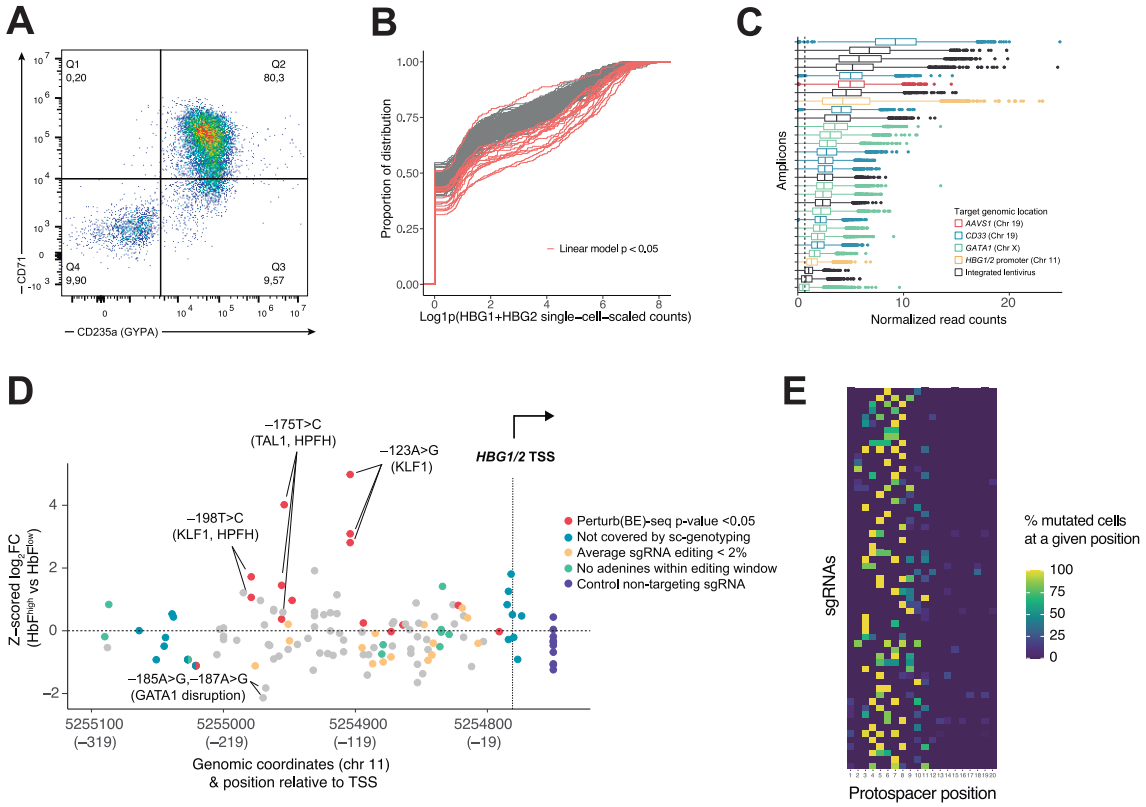
(C) Scatterplot of the counts of the top sgRNA (measured using CROP-seq transcript counts detected directly on scRNA-seq reads, STAR Methods) in each cell relative to the counts of all sgRNAs detected in that cell. The dashed line denotes the separation between cells with a dominant perturbation detected (left) and cells with non-dominant perturbations (which might be caused by doublets).

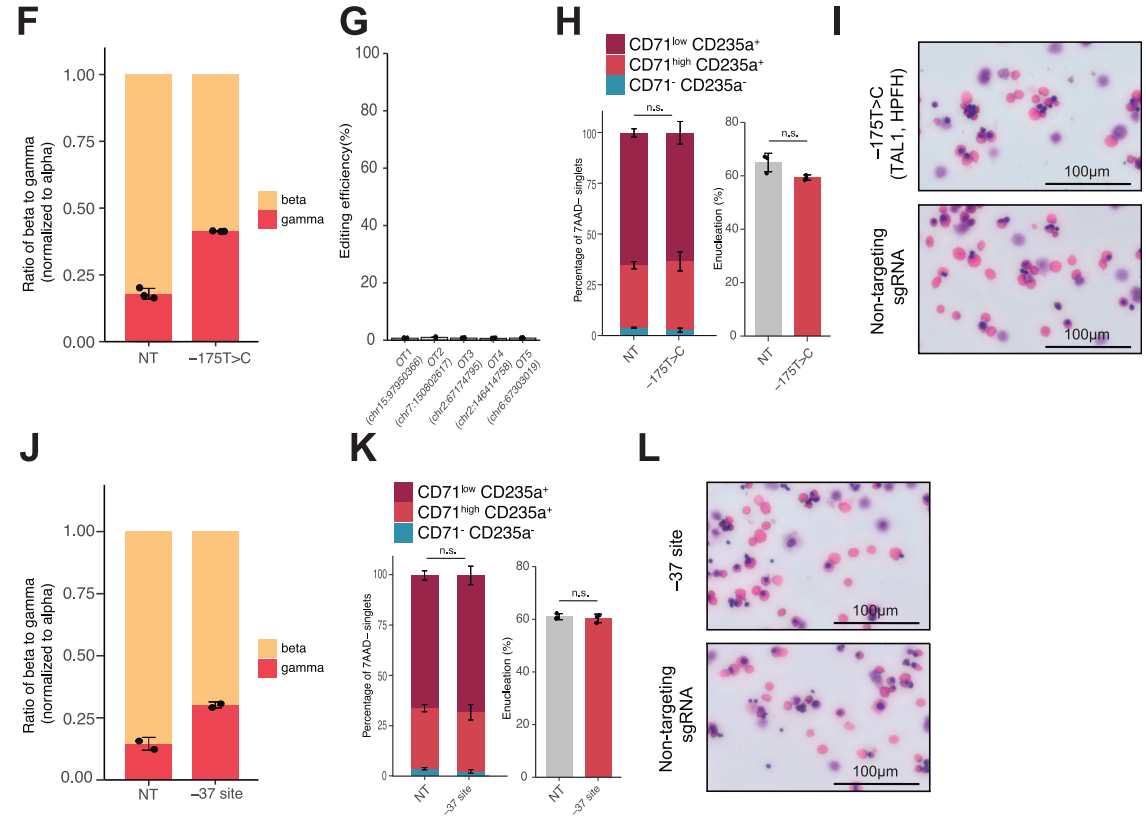
(D) Percentage of cells with a detected CROP-seq transcript and with a discernable dominant sgRNA in the scRNA-seq library sequenced at an average of 20,000 reads per cell and in the enrichment PCR library, respectively.

(E) Scatter plot of the total sgRNA counts per cell directly detected in scRNA-seq reads of CROP-seq transcripts and those detected using enrichment PCR. The dashed line has a slope of 1.

(legend continued on next page)

(F) Area-proportional Venn diagram of the perturbation assignment in cells with at least 1 CROP-seq transcript count directly detected in scRNA-seq reads and in enrichment PCR. Perturbation identity was determined as the perturbation with the highest number of CROP-seq counts in a cell.
(G) Distribution of VEX-GFP fluorescence levels (the positive selection marker expressed in our modified CROP-seq vector) across HSPCs and derived multi-lineages following initial infection at low MOI and prior to Perturb(BE)-seq. Both groups were downsampled to the same number of cells for comparison.
(H) UMAP plots displaying representative markers of each of the lineages, as well as CD33 RNA expression.
(I) Violin plots displaying the absolute CD33 antibody counts by cluster.
(J) Cumulative distribution of single-cell CD33 protein counts (scaled by cluster average) across cells transduced with CD33 splice donor 1 sgRNA vector and electroporated with ABE and cells transduced with a non-targeting (NT) sgRNA vector and electroporated with ABE. Each individual distribution represents one of the 8 clusters in Figure 3B.





(legend on next page)

Figure S4. Identification of nucleotides modulating fetal hemoglobin levels in primary hematopoiesis and validation of sites, related to Figure 4

(A) Dot plot of maturing human hematopoietic stem/progenitor cell (HSPC)-derived erythroblasts (CD71+CD235a+) on day 13 of differentiation. HSPCs were transduced with the *HBG1/2* library and electroporated.

(B) Cumulative distribution of the log1p(*HBG1* + *HBG2* single-cell scaled counts) for each sgRNA in the screen. The red lines highlight significant sgRNAs in the linear model shown in Figure 4B and discussed in STAR Methods.

(C) Normalized read counts for each of the 30 amplicons used to genotype single HSPCs transduced with lentiviral libraries and base-edited. The dashed line represents 20% of the mean reads per amplicon and per cell.

(D) *Z* scored $\log_2$(FC) for each sgRNA between HbF^{high} and HbF^{low} erythroblasts from the functional screen.

(E) Percentage of single cells with edits at each position of the protospacer for mutated cells from each sgRNA from the pooled single-cell genotyping experiment, for sgRNAs with sufficient coverage.

(F) Quantitative real-time PCR measurements of *HBG1/2* and *HBB* transcripts (normalized by the levels of *HBA2* transcripts) in HSPCs treated with ABE precomplexed with an sgRNA targeting the −175T>C mutation and HSPCs with ABE precomplexed with a non-targeting sgRNA, which were subsequently differentiated into the erythroid lineage (day 15). Error bars represent the standard deviation and dots represent 3 independent electroporations.

(G) Editing efficiencies at the 5 predicted off-target locations with 1 and 2 mismatches in the protospacer from the sgRNA targeting −175T>C using CRISPR-OFFinder.[120]

(H) Erythroid cells differentiated from HSPCs treated with ABE precomplexed with an sgRNA targeting the −175T>C mutation and HSPCs with ABE precomplexed with a non-targeting sgRNA (day 20). Left, percentage of mature terminal erythroid CD71− CD235a+, late erythroid precursors CD71+ CD235a+ and non-erythroid CD71− CD235a− cells for each condition, from 7AAD− singlets. Right, percentage of enucleated cells (7AAD/Hoechst negative) from the $GYPA^{high}$ population. Data from 3 independent electroporations. Two-tailed unpaired t test.

(I) Representative microscopy images from erythroid cells differentiated (day 18) from HSPCs treated with ABE precomplexed with an sgRNA targeting the −175T>C mutation and HSPCs with ABE precomplexed with a non-targeting sgRNA. May-Grünwald-Giemsa stain.

(J) Quantitative real-time PCR measurements of *HBG1/2* and *HBB* transcripts (normalized by the levels of *HBA2* transcripts) in HSPCs treated with ABE precomplexed with an sgRNA targeting the −37 region and HSPCs with ABE precomplexed with a non-targeting sgRNA, which were subsequently differentiated into the erythroid lineage (day 15). Error bars represent the standard deviation and dots represent 2 independent electroporations.

(K) Erythroid cells differentiated from HSPCs treated with ABE precomplexed with an sgRNA targeting the −37 site and HSPCs with ABE precomplexed with a non-targeting sgRNA (day 20). Left, percentage of mature terminal erythroid CD71− CD235a+, late erythroid precursors CD71+ CD235a+ and non-erythroid CD71− CD235a− cells for each condition, from 7AAD− singlets. Right, percentage of enucleated cells (7AAD/Hoechst negative) from the $GYPA^{high}$ population. Two-tailed unpaired t test.

(L) Representative microscopy images from erythroid cells differentiated (day 18) from HSPCs treated with ABE precomplexed with an sgRNA targeting the −37 site and HSPCs with ABE precomplexed with a non-targeting sgRNA. May-Grünwald-Giemsa stain.

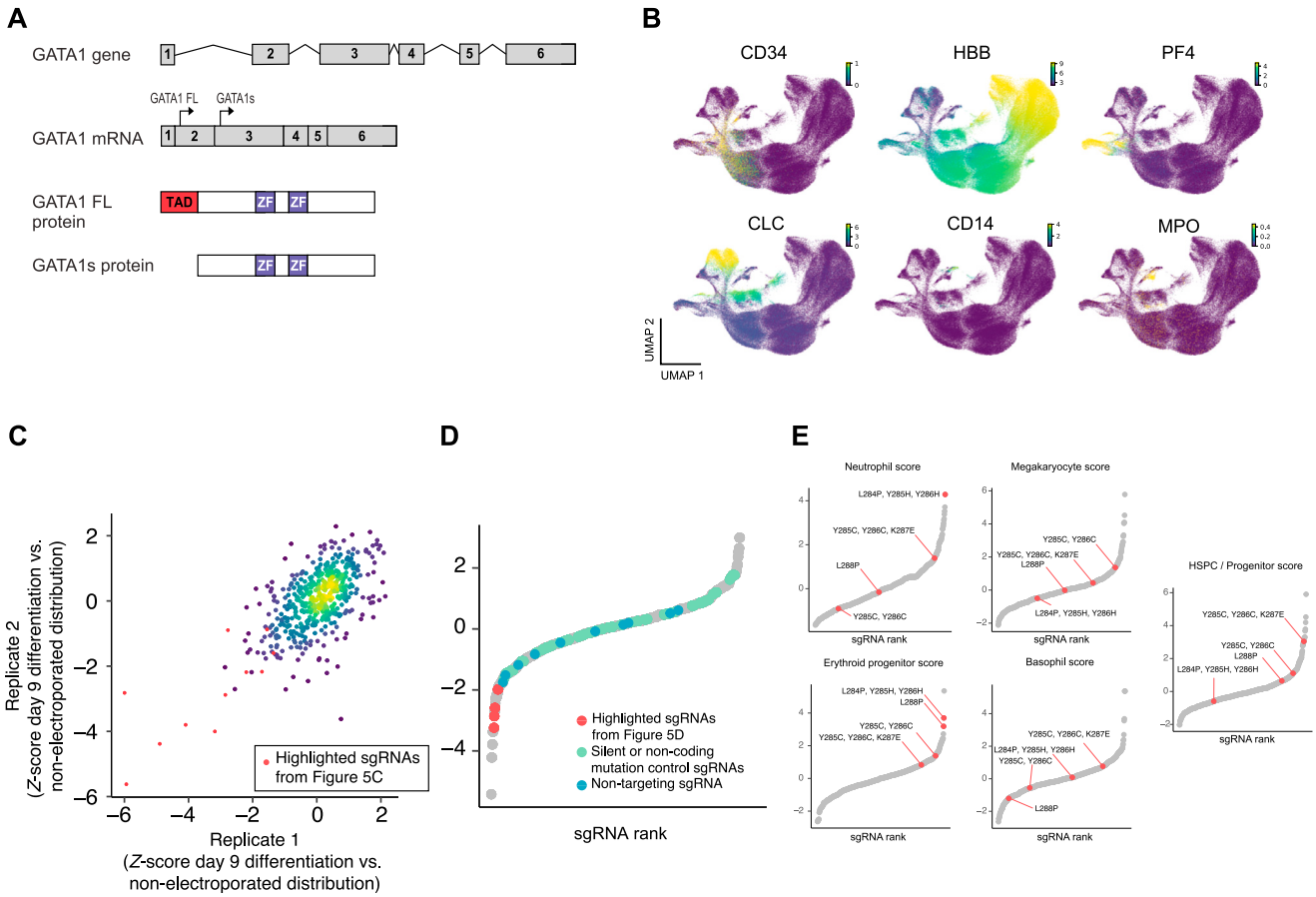


Figure S5. Systematic mutagenesis of the master hematopoietic transcription factor GATA1 with Perturb(BE)-seq and pooled single-cell screen genotyping, related to Figure 5

(A) Schematic of the *GATA1* gene, the *GATA1* transcript, and the full length (GATA1 FL) and short (GATA1s) isoforms it gives rise to using its two start codons.[75] TAD, transactivation domain; ZF, zinc finger.

(B) UMAP plots displaying representative markers of each of the lineages.

(C) *Z* score of the $\log_2$(FC) of sgRNAs between the day 9 differentiation time point and the transduced, non-electroporated HSPC control cells using bulk amplicon sequencing. Data are shown for two biological replicates.

(D) Same as Figure 5D (*Z* scored ratio between the number of cells in erythroid lineages and the number in non-erythroid lineages for each sgRNA –erythroid score–), but with non-targeting control sgRNAs and sgRNAs targeting silent or non-coding control mutations also highlighted. Screen hits in close proximity on the GATA1 C-terminal zinc finger from Figure 5D are also shown.

(E) Lineage-specific enrichment scores, calculated analogously to the erythroid score (Figures 5 and S5D). Hits clustering on the C-terminal zinc finger (presented in Figure 5D) are highlighted, underscoring their specificity to the erythroid lineage.

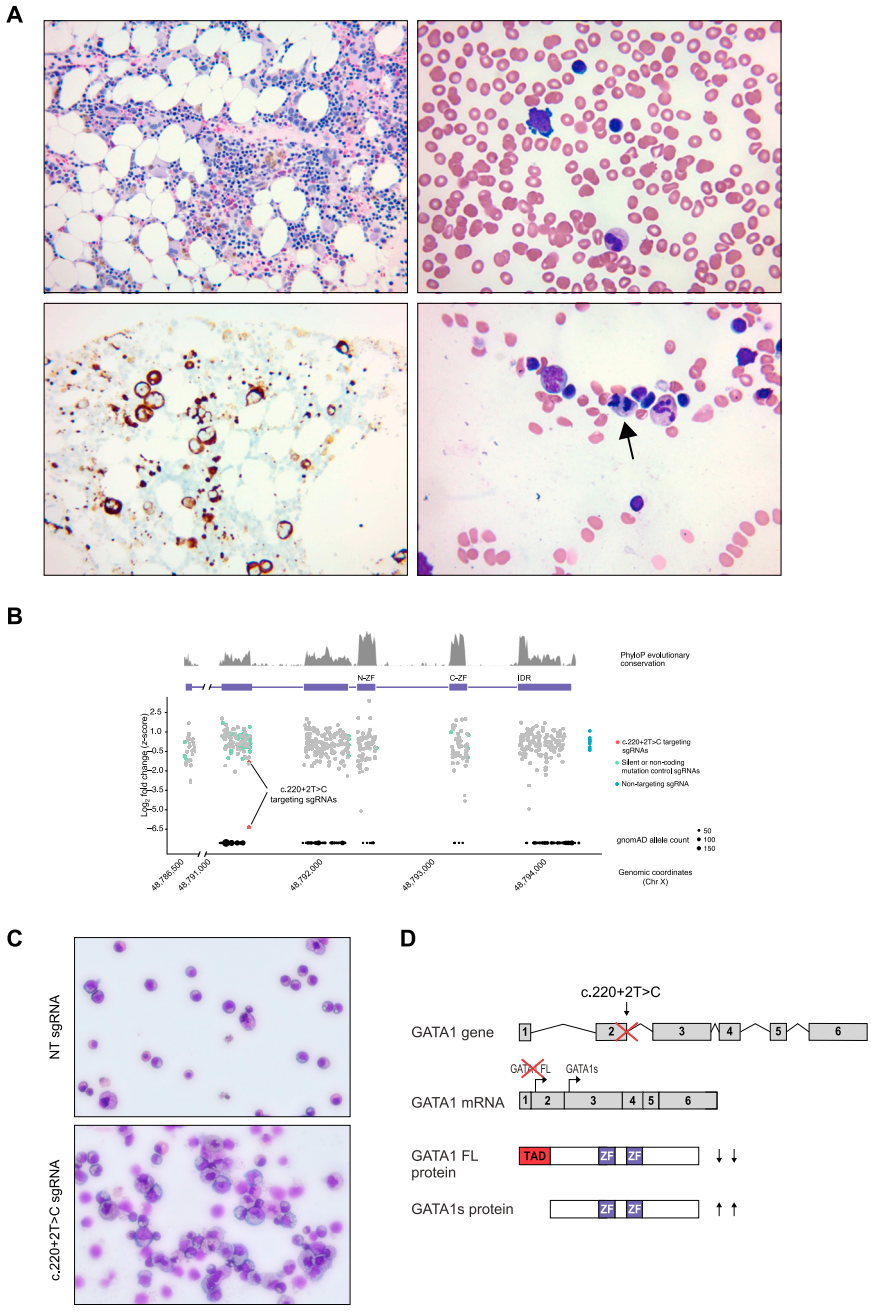


Figure S6. Defining the pathogenicity of a GATA1 variant of unknown significance through base-editing screens, related to Figure 6

(A) Representative bone marrow aspirates of the patient with the *GATA1* c.220+2T>C variant of unknown significance. Top left, erythroid hypoplasia with dysplastic megakaryocytic lineage showcasing hypolobated forms. Top right, subset of erythroid cells with irregular nuclear borders and cytoplasmic/nuclear asynchrony. Bottom left, immunohistochemistry staining of CD42b, a megakaryocyte marker, which also highlights hypolobated morphology. Bottom right, the arrow points to a binucleated erythroid cell, albeit present at very low frequency in smears.

(B) Same as Figure 5C (*Z* scored $\log_2$(FC) of sgRNA in cells sampled on day 9 of erythroid differentiation with respect to transduced cells prior to electroporation in the *GATA1* screen using bulk amplicon sequencing), but with the two sgRNAs targeting c.220+2T>C highlighted. Control sgRNAs that were included to target silent mutations or non-coding regions and non-targeting controls are also shown. PhyloP evolutionary conservation scores and gnomAD allele counts at each position are highlighted for reference (STAR Methods).

(C) Representative microscopy images from cells differentiated from HSPCs treated with ABE precomplexed with sgRNA 1 targeting the c.220+2T>C mutation and HSPCs with ABE precomplexed with a non-targeting sgRNA. A reduction in the presence of erythroblasts with a dominance of myeloid cells and precursors is observed. Day 10 post-electroporation. May-Grünwald-Giemsa stain.

(legend continued on next page)

(D) Schematic of the pathogenic mechanism of the 220+2T>C *GATA1* mutation. The 220+2T>C mutation results in preferential splicing of *GATA1* to the short (GATA1s), rather than full-length (GATA1 FL) isoform, thereby perturbing erythropoiesis and causing hypoplastic anemia. TAD, transactivation domain; ZF, zinc finger.

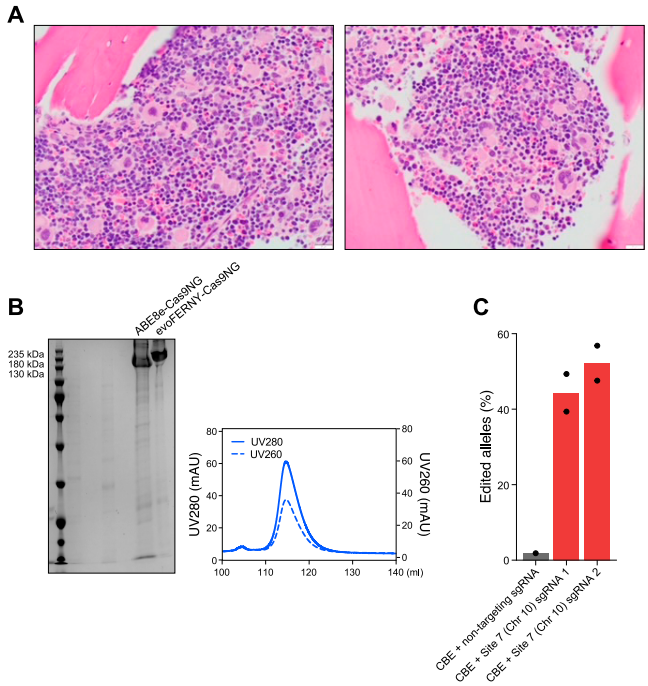


Figure S7. Using cytosine base editors to characterize a variant of unknown significance and to expand the number of targetable variants in screens, related to Figure 7

(A) Representative bone marrow aspirates of a male patient with a GATA1 VUS (c.218C>T) in the vicinity of the second exon-intron junction. Bone marrow biopsy revealed hypercellularity with atypical megakaryocytic hyperplasia, erythroid hypoplasia, and some dyserythropoietic features.

(B) Left, representative polyacrylamide gel electrophoresis of purified cytosine base editor (evoFERNY-Cas9NG). ABE8e-Cas9NG is shown for comparison. Right, quantification of the UV280 and UV260 ratios for evoFERNY-Cas9NG.

(C) Percentage of edited alleles using evoFERNY-Cas9NG precomplexed with two different sgRNAs targeting site 7 (chromosome 10). evoFERNY-Cas9NG precomplexed with a non-targeting sgRNA is shown as control.