BAR-Seq Clonal Tracking of Gene Edited Cells

Samuele Ferrari[1,2,*,#], Stefano Beretta[1,*], Aurelien Jacob[1,3], Davide Cittaro[4], Luisa Albano[1], Ivan Merelli[1,5,†], Luigi Naldini[1,2,†], Pietro Genovese[1,†].

[1]San Raffaele Telethon Institute for Gene Therapy (SR-Tiget), IRCCS San Raffaele Scientific Institute, 20132 Milan (MI), Italy.

[2]Vita-Salute San Raffaele University, 20132 Milan (MI), Italy.

[3]Milano-Bicocca University, 20900 Monza (MB), Italy.

[4]Center for Omics Sciences, IRCCS San Raffaele Scientific Institute, 20132 Milan (MI), Italy.

[5]National Research Council, Institute for Biomedical Technologies, 20090 Segrate (MI), Italy.

[*]Equally contributing Authors.

[†]These Authors share senior authorship.

PG current address is: Gene Therapy Program, Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Department of Pediatric Oncology, Harvard Medical School, Boston, MA, USA.

[#] Corresponding author: ferrari.samuele@hsr.it

RELATED LINKS

Key references using this protocol

Ferrari S, Jacob A, Beretta S, Unali G, Albano L, Vavassori V, Cittaro D, Lazarevic D, Brombin C, Cugnata F, Kajaste-Rudnitski A, Merelli I, Genovese P, Naldini L. *Efficient gene editing of human long-term hematopoietic stem cells validated by clonal tracking.* Nat Biotechnol. 2020 Nov;38(11):1298-1308. doi: 10.1038/s41587-020-0551-y. Epub 2020 Jun 29.

ABSTRACT

Gene editing by engineered nucleases has revolutionized the field of gene therapy by enabling targeted and precise modification of the genome. However, the limited availability of methods for clonal tracking of edited cells resulted in paucity of information on diversity, abundance and behavior of the engineered clones. Here, we detail the wet and bioinformatic BAR-Seq pipeline, a barcoding-based strategy for clonal tracking of cells harboring homology-directed targeted integration. We present the BAR-Seq web application, an online freely available and easy-to-use software which allows to perform clonal tracking analyses on raw sequencing data without any computational

resources or advanced bioinformatic skills. BAR-Seq can be applied to most editing strategies and we describe its use to investigate clonal dynamics of human edited hematopoietic stem/progenitor cells in xenotransplanted hosts. Notably, BAR-Seq may be applied both in basic and translational research contexts to investigate the biology of edited cells and stringently compare editing protocols at clonal level. Our BAR-Seq pipeline allows library preparation and validation in few days and clonal analyses of edited cell populations in one week.

INTRODUCTION

Viral vectors are widely exploited to transfer genetic information into cells of interest. Gene therapy takes advantage of viral vectors to introduce therapeutic transgenes in patients' cells and, therefore, holds great promise for the treatment of several diseases[1]. The semi-random genomic integration of some viral vectors, such as gamma-retroviral and lentiviral vectors (LV), enables univocal and permanent marking of genetically modified cells and their progeny, thus identifying cell clones which can be tracked over time and through space by means of vector integration sites[2]. Clonal tracking of genetically modified cells in preclinical and clinical studies expanded the knowledge on safety and effectiveness of gene therapy, as well as giving remarkable insights on target cell biology and differentiation[3–7].

In the last years, gene editing has broadened the scope and means of genetic manipulation by allowing precise integration in pre-selected safe genomic loci or *in situ* functional correction of a mutant gene[8]. Engineered nucleases, such as CRISPR/Cas, are used to induce a DNA double strand break (DSB) at the locus of interest, which can then be repaired by either the non-homologous end joining (NHEJ) or the homology-directed repair (HDR) cellular machinery (**Fig. 1a**). The NHEJ pathway rejoins the free DNA ends of the break while often introducing small base insertion or deletions (indels), thus leaving a permanent genetic scar at the edited locus. Conversely, high-fidelity HDR pathway can repair the DSB using a DNA template bearing homology to the target site.

Clonal tracking of edited cells can provide relevant information on the complexity and dynamics of edited cell clones, thus expanding the characterization of the engineered cell product and guiding its development and optimization in view of perspective clinical translation. Genomic scars introduced by NHEJ might provide a surrogate clonal tracker in applications aiming at gene disruption, albeit underpowered by the recurring generation of few dominant indels. On the other hand, gene correction or targeted integration strategies based on HDR lack such possibility, as all editing events will have the same sequence. Recently, we developed a barcoding-based strategy (BAR-Seq) that allows clonal tracking of edited cells by means of unique molecular identifiers (barcodes, BARs) embedded in the DNA template for HDR[9]. Molecular barcoding is a well-

established and successful method for lineage and cell tracking[7,10,11]. BAR-Seq is a versatile and portable "three-steps" clonal tracking pipeline (**Fig. 1b)** based on:

i) generation of barcoded HDR template libraries;

ii) editing of the locus of interest in the selected target cell population (here hematopoietic stem/progenitor cells, HSPCs) using the editing nuclease of choice and the barcoded template library;

iii) deep sequencing of BAR-Seq amplicons from the edited cells or their progeny at different times post-treatment and bioinformatic analyses for the retrieval of BAR sequences and their abundances.

Here we describe the BAR-Seq workflow applied to human HSPC gene editing and report a robust experimental and bioinformatic pipeline to assess clonal composition of edited cells. BAR-Seq enables: i) characterization of the *in vivo* repopulation capacity of gene edited human HSPCs transplanted in murine recipients; ii) validation of improved editing protocols; and iii) identification of experimental conditions preserving broader clonal repertoire of edited cells in recipient hematopoiesis.

Additionally, we provide a flexible, freely available and user-friendly web application (http://www.bioinfotiget.it/barseq), which eases and speeds up clonal tracking analyses of gene edited cells from raw fastq sequencing data. Of note, this web application and its bioinformatic pipeline are suitable for clonal tracking analyses based on viral barcoded vectors, thus providing a useful tool for viral or gene editing clonal tracking also by researchers with limited programming experience.

APPLICATIONS OF THE METHOD

BAR-Seq can conceivably be applied to any cell type of interest from any eukaryotic species. Templated sequence editing will be more efficient in actively cycling cells, since HDR is restricted to S/G2 phase of the cell cycle[12]. Quiescent and slowly cycling cells, such as HSPCs, are less permissive to HDR[13]. We did not observe any detrimental consequence of introducing BAR in the HDR template, neither on HDR efficiency nor on cellular response to the editing procedure[9].

BAR-Seq can be applied to bulk *ex vivo* cultured cells and used to track engraftment and lineage output upon transplantation. The high sensitivity of the barcoding-based platform allows to assess clonal abundance and distribution even within rare sorted cell subpopulations, such as hematopoietic progenitors.

BAR-Seq is expected to be compatible with any engineered nuclease, including Zinc Finger Nucleases (ZFN), Transcriptional activator-like effector nucleases (TALEN) and CRISPR/Cas

variants. In our studies, we successfully used ZFN (unpublished data) and CRISPR/Cas9[9] as nucleases.

Barcodes can be embedded in HDR templates designed to target virtually any genomic locus, both in somatic[9] and sex-linked chromosomes (unpublished data), albeit some considerations should be made in the former case (see "Limitations"). BAR-Seq could be adapted to different delivery vehicles of HDR templates. Therefore plasmid, double stranded (ds)DNA, single stranded (ss)DNA, integrase-defective lentiviral vector (IDLV) and adeno-associated vector (AAV) can be used for BAR-Seq experiments. In our studies, we successfully used AAV serotype 6 (AAV6)[9], which has been shown to efficiently transduce hematopoietic cells[14], and VSV.G pseudotyped IDLV (unpublished data). Barcodes can conceivably be included in any part of the HDR template, either in transcribed or non-transcribed regions. The use of transcribed barcodes (eBAR-Seq) would allow powerful combination of single-cell transcriptomic studies with lineage tracking information, as previously reported in other studies based on barcoded LV[15].

Finally, BAR-Seq bioinformatic pipeline also identifies the most abundant ("dominant") cell clones in the HDR-edited population, thus focusing sharing analyses on those clones more robustly contributing to the *in vitro* outgrowth or to host repopulation. In our experimental settings, this approach allowed to study the clonal behavior of long-term repopulating HSPC clones and removed the background signal derived from short-lived cells providing limited output. Of note, the frequency of dominant HSPC clones long-term after transplant in murine recipients, as calculated by BAR-Seq, well fitted with estimates in limiting-dilution repopulation assays[16,17], thus validating our clonal tracking pipeline. Of note, clonal analysis of the most abundant clones may find its application beyond hematopoiesis, including most setting based on host repopulation and modelling of clonal expansion.

COMPARISON WITH OTHER METHODS

In the gene editing field, fluorescent reporter genes (e.g. GFP) have been extensively used to provide a surrogate and easy readout to track and quantify *bona-fide* HDR-edited cells and their progeny[14,18,19]. However, this approach does not allow to discriminate among different edited clones. BAR-Seq enables clonal tracking of HDR-edited cells at single-cell resolution, even in absence of reporter-expressing cassettes, by offering the possibility to investigate proliferation, differentiation, self-renewal and long-term maintenance of HDR-edited cell clones, as well as to stringently compare different editing protocols and reagents.

Genomic scars (i.e. indels) introduced by NHEJ at the nuclease target site can be exploited as markers of clonality and for lineage tracking[20,21] by measuring indel diversity. Of note, the clonal

diversity might be underestimated by biased insertion/deletion of nucleotides[22] or restoration of the wildtype (WT) sequence. Furthermore, NHEJ or microhomology-mediated "large" deletions might result in dropout of cell clones due to the loss of primer binding sites on the target region[23]. Finally, this approach cannot interrogate HDR-edited cells.

EXPERIMENTAL DESIGN

Design and production of the barcoded plasmid library. The conventional (non-barcoded) template for HDR is composed of a transgene or therapeutic cassette framed by two sequences bearing homologies for the intended target site (**Fig. 2a, left**). The maximal length of the HDR template is dependent on the cargo capacity of the vector used for its delivery, which is approximately 4.7 and 9 kilobases (kb) for AAV and IDLV, respectively. Protocols describing the optimal configuration of conventional HDR template have been previously published[24–26]. The non-barcoded template for HDR can be designed and purchased by gene synthesis services and then subcloned into the appropriate backbone (e.g. any transfectable plasmid, pAAV-MCS or pCCL-LV transfer plasmids) depending on the vehicle chosen for template delivery. Here, we highlight the key points for the generation of plasmid libraries carrying the BAR-Seq HDR template.

First, we advise to generate the conventional HDR template in a backbone suitable for the final purpose. For AAV, the whole HDR template must be cloned between inverted terminal repeats (ITRs), conventionally derived from AAV2. Alternatively, the HDR template can be cloned in a third-generation self-inactivating transfer construct suitable for IDLV production or in any transfectable plasmid DNA. In any case, two unique restriction sites generating incompatible ends (such as SphI and Bsu36I) spaced by >10 nucleotides (nts) should be included at the intended site for BAR cloning to avoid BAR concatemers and facilitate the generation of the BAR-Seq HDR template library. BAR must be cloned either inside (transcribed eBAR) or outside (genomic BAR) an expressed cassette and always between homology arms to ensure its incorporation into the genome upon HDR (**Fig. 2a, right**). BAR length and consensus sequence must be carefully evaluated in order to reach an adequate library complexity (i.e. number of unique molecular identifier) and minimize biases in BAR structure. BAR-Seq requires forward and reverse primers binding upstream and downstream BAR cloning site. To exclusively amplify the on-target BARs, one of the two primers must bind the genomic region outside the homology arm ("In-Out" PCR approach) (**Fig. 2b**).

The maximal theoretical complexity of the library having length x can be calculated as 4^x, assuming all positions can include all nucleotides. More in general, the complexity is calculated as $C = \prod D_i$, where D_i is the number of allowed nucleotides in position i, and the effective BAR length is $\log_4(C)$. The decision about BAR length strictly depends on the expected clonality of the cell

population of interest. Indeed, the longer is the BAR, the more complex will be the library and the lower will be the probability of tagging >1 HDR-edited cell with the same BAR[2,27]. To minimize the chance of having two individual cell clones tagged with the same BAR, we recommend an effective BAR length >12 nucleotides and the library complexity >10^3-fold higher than the expected complexity of the HDR-edited cell population. Indeed, the risk of having two cells sharing the same barcode (collision) follows a binomial distribution. However, this distribution can be approximated by a Poisson distribution with $\lambda = n/b$, where λ is the inverse of the aforementioned threshold (10^{-3}), when the number of cells (n) and barcodes (b) is sufficiently high, as in this protocol and other barcoding techniques. Therefore, the probability of uniquely barcoded cells is $P_{\text{Poisson}}(0; \lambda)$, which is >99% already when $\lambda = 10^{-2}$, corresponding to a 1:100 proportion (where P_{Poisson} is the probability mass function of the Poisson distribution) (**Supplementary Table 1**). Previous works indicate that even 1:100 proportion[28] or 1:50 proportion[11] are sufficient to minimize collision events. Since the ratio is calculated on the theoretical complexity, we choose a conservative rule of thumb to take into account the subsequent reduction of complexity due to the efficiency of the library preparation step. In the BAR structure, some positions may be fixed or limited to few bases to avoid generation of BARs carrying the cloning restriction sites, which would be preferentially ligated, and thus overrepresented, in the library because of shorter length[27]. For this purpose, we designed the BAR consensus sequence as 5'-NNYNNNTNTNNNNRTNDNNNHH-3', having a maximal theoretical complexity of ~2.899 x 10^{10} (instead of ~1.759 x 10^{23} with Ns only). Of note, the real complexity of the library will be lower than the theoretical one due to bottlenecks in cloning and plasmid/viral preparation. With our protocol, we easily obtained final library complexity from 3 x 10^5 to 8 x 10^5 unique BARs with homogeneous representation. Processing of BARs involves a step of collapse of sequences closer than a given edit distance; this process is an additional contributor to collision events. The library complexity sets an upper bound on the number of cells that are needed to generate collisions. The distribution of edit distances in a random set of sequences is modeled after a Gumbel distribution. We estimated the parameters for different barcode lengths and the probability of collisions at $d \leq 3$. The number of sequences, hence the number of edited cells, needed to generate a collision is estimated as in the "Birthday attack" problem and is tabulated in the **Supplementary Table 2** for $d \in \{1, 2, 3\}$.

Regarding the preparation of the BAR-Seq plasmid library, a synthetized single-stranded oligonucleotide (ssODN) library embedding the BAR sequence is amplified by few PCR cycles to generate the complementary strand, digested with the appropriate restriction enzymes and subcloned as insert in the non-barcoded plasmid (**Fig. 2c**). In the single-stranded oligonucleotide, we suggest including stuffer sequences flanking the restriction sites in order to verify for successful digestion of

the amplified product. Plasmid amplification is performed by ultra-efficient chemical transformation in recombinase-negative *Escherichia coli*, which are plated on lysogeny broth (LB) agar plates. In parallel, transformation of equimolar amount of dephosphorylated digested backbone ligated in absence of the barcoded insert should be performed as control. The number of colonies in the control plate should be 10^4-fold lower compared to the one in the other plates in order to have <1 out of 10^4 HDR-edited clones being untraceable in the cell population. An estimation of the total number of colonies provides a useful indication of the library complexity. Colonies are collected and mixed by scraping, and bacteria are grown in LB at 30°C to minimize recombinogenic events. The BAR-Seq plasmid library is purified from the bacteria outgrowth. Of note, several parameters during BAR-Seq library cloning might affect the number of unique BARs and, therefore, can be modified to obtain final libraries with higher complexity depending on the experimental needs. In particular, the number of PCR cycles on the ssODN library can be decreased to improve library diversity. The amount of ligated plasmid, the number of ligation reactions and the number of transformed *E. coli* can be scaled up to achieve BAR-Seq libraries of higher complexity. Electrocompetent rather than chemically competent recombinase-negative cells can be used to further increase final library complexity. High-throughput sequencing of the BAR-Seq plasmid library is highly recommended to assess its diversity (i.e. high complexity and equal representation of the different BARs) before moving to the next steps of the protocol (see also **Box 1** and **BAR-Seq bioinformatic analyses**).

Alternatively, generation of the BAR-Seq plasmid library could be also achieved by Gibson assembly[29], albeit insert >100 bp should be used to maximize its efficiency. Other detailed protocols have been proposed for the generation of barcoded plasmid libraries[2].

Barcoded viral library production, purification and titration. The choice of the delivery vehicle for the HDR template is pivotal to maximize efficiency and tolerability of gene editing, and strictly depends on the target cell type and the specific application. Non-viral delivery methods, such as injection, lipofection or nucleofection of plasmids, single- or double-stranded DNA[30,31], and viral vector transduction, such as IDLV[32] or AAV[26], can be used to deliver HDR templates into mammalian cells. In our studies, highly efficient gene editing in human HSPCs was achieved by AAV6, which is the most efficient AAV serotype for HSPC transduction[14]. BAR-Seq libraries are also compatible with other serotypes depending on the cell types of interest. Briefly, AAV are produced in HEK293 adherent mammalian cells by co-transfection of two or three plasmids containing (i) the AAV genes (i.e. *rep* and *cap*), (ii) the essential adenoviral genes VA, E2A and E4, and (iii) the AAV genome with a maximal size of 4.7 kb framed by ITRs required for genome replication and encapsidation[33]. Plasmids required for AAV production are commercially available (e.g. the AAV genome by Agilent

Technologies and the pDGM plasmid for rep-cap and helper expression by Addgene/Russell's lab). Detailed protocol for AAV production, purification and titration, as well as further indications about AAV6 production and their use in *ex vivo* gene editing experiments have been previously published[26,34,35]. Alternatively, custom AAV can be produced by specialized companies. In general, the experimental workflow and the reagents for production and purification of viral libraries are identical to those of conventional viral vectors. The number of transfected cells during viral library preparation is the most critical parameter to avoid significant loss of library complexity compared to the plasmid library; as a rule of thumb, transfection of $1.1x10^9$ HEK293 cells suffices for the production of an AAV library starting from a plasmid library with 10^5-10^6 unique BARs[35]. Scaling up of the AAV production may be necessary for more complex libraries. In any case, diversity assessment of the BAR-Seq viral library by transduction of highly permissive cell lines is highly recommended before moving on to gene editing experiments (see also **Box 2** and **BAR-Seq bioinformatic analyses**).

Gene editing procedure. Gene editing protocol varies according to the application and the target cell type. Despite BAR-Seq could be also applied to *in vivo* gene editing, its most straightforward application is in the context of *ex vivo* gene editing for clonal tracking of HDR-edited cells capable of host engraftment. Protocol for the design of CRISPR/Cas gene editing strategies has been previously published[36]. Here, we briefly report our optimized gene editing procedure for human HSPCs as described in[9] and suitable for BAR-Seq analyses. HSPCs can be collected from different donor sources, such as cord blood (CB), mobilized peripheral blood (mPB) or bone marrow upon informed consent and in compliance with protocols approved by the relevant Institutional Review Boards. Human HSPCs from these sources can be also purchased from different sellers (e.g. Lonza, STEMCELL Technologies). To favor HDR and maintain long-term repopulating potential, HSPCs are stimulated in culture with early acting cytokines (SCF, FLT3L, TPO, IL6) in the presence of the stem-cell preserving compounds StemRegenin 1 (SR1)[37] and UM171[38,39]. SR1 and UM171 allow more robust hematopoietic output from edited CB HSPCs in hematochimeric mouse models and moderately increase clonality of short-term engrafting progenitors[9]. After 3 days of stimulation, CRISPR/Cas nucleases are delivered by nucleofection as ribonucleoproteins (RNPs) composed by the purified Cas protein and the single guide RNA (sgRNA) synthetized with chemical modifications to stabilize its structure and avoid innate cellular responses impacting on cell biology[9,19]. Alternatively, ZFNs, TALENs and CRISPR/Cas can be delivered as HPLC purified mRNA[19,39–41]. Nucleofection can be also exploited to co-deliver barcoded HDR templates, such as dsDNA[42] or ssDNA[43], or editing enhancers[9]. When using AAV6 for the delivery of barcoded HDR template,

HSPCs are transduced immediately after nucleofection[14]. Alternative protocols for AAV6 or IDLV-based gene editing in HSPCs were previously described[9,26,44]. Human edited HSPCs can then be transplanted by tail-vein injection in immunodeficient mice (NSG or NSGW41) to evaluate their long-term repopulating and self-renewal potential. We suggest transplanting the same number of culture-initiating HSPCs/mouse (i.e. the outgrowth of the same number of starting cells at the beginning of the culture) across experimental conditions, rather than the same number of HSPCs after editing. This procedure allows to stringently compare the impact of different editing treatments on HSPC repopulation capacity. Of note, transplantation of high number culture-initiating HSPCs/mouse may lead to saturation of the hematopoietic niche[9] and may camouflage differences of HSPC reconstitution capacity across experimental conditions. We suggest transplanting 1.0-1.5x10^5 and <5x10^5 culture-initiating cells/mouse when editing CB or mPB HSPCs, respectively.

Quantification of the editing efficiency and phenotypic characterization of the edited cell population (either *ex vivo* or *in vivo*) might be highly relevant to complement and interpret BAR-Seq data. If reporter genes (e.g. NGFR, GFP) are embedded in the barcoded HDR template, the percentage of reporter-expressing cells can be used as readout of the fraction of cells harboring integration. For more reliable quantification it is advisable to measure reporter expression in the treated cells after several days of culture, when multiple rounds of proliferation have diluted any residual episomal HDR template, which may otherwise contribute to reporter expression and confound the assessment of integrated copies. To measure HDR efficiency at molecular level, we perform digital droplet PCR (ddPCR)-based assays that quantify the copies of edited alleles and those of a reference unedited gene[39,42]. We advise to design the ddPCR amplicon following an "In-Out" PCR approach in order to specifically amplify the donor-genome junction (either 5' or 3') upon on-target HDR-mediated integration (**Fig. 2b**). Although these assays must be designed and optimized for each editing strategy, their high sensitivity and precision allows reliable quantification of HDR editing events. Notably, the ddPCR amplicon can be designed to overlap with the BAR-Seq amplicon, whenever possible. If clonal tracking studies are also extended to NHEJ-edited cells, accurate quantification of the indels frequency is directly provided by targeted NGS of the nuclease target site[45]. Alternatively, mismatch-sensitive endonuclease assay or Sanger sequencing followed by deconvolution analysis (e.g. TIDE)[46] may be performed to assess the overall nuclease cutting efficiency. Independently of the method used for the assessment of editing efficiency, we recommend including adequate controls in the experimental design (i.e. HDR donor only, untreated cells) and performing HDR-editing analyses >3 days after editing procedure, if applicable, to minimize any confounding effect due to the presence of episomal HDR template (e.g. PCR jumping).

BAR-Seq amplicon design and library preparation. BARs are retrieved from the BAR-Seq plasmid/viral library, when assessing its complexity and diversity, or genomic DNA (gDNA) of edited cells.

In the first case, BAR sequences can be extracted by PCR amplification using primers flanking the BAR. To ensure an adequate coverage of the original library, the number of sequencing reads should be set at least 10-fold higher than the expected library complexity, as estimated by the total number of bacterial colonies counted. Amplicons can be sequenced by single-end Illumina MiSeq (MiSeq™ Reagent Kit v3), NextSeq or HiSeq platforms, depending on the required number of reads. Consecutive sequencing rounds of the same library can be performed to increase sequencing depth when an insufficient number of reads has been obtained.

In the second case, the number of cells used for the analysis depends on the expected complexity of the population of interest and the percentage of HDR-edited cells: the higher is the expected population complexity, the higher is the number of HDR-edited cells to be harvested and analyzed by BAR-Seq to exhaustively investigate cell population complexity. From another perspective, the number of HDR-edited cells analyzed determines the abundance of the rarest BAR (i.e. cell clone) that can be identified. As a rule of thumb, clones representing 0.01% of an HDR-edited cell fraction, whose proportion is 10% of the bulk population, can be recovered only if analyzing at least 100,000 bulk cells. As also indicated in other clonal tracking pipelines[2], we suggest performing BAR-Seq analysis on 10-fold higher number of bulk cells to ensure better results. Importantly, sorting of reporter-expressing cells, if applicable, as *bona-fide* HDR-edited cells is not required since the BAR-Seq amplicon design allows to extract BARs from a bulk population.

In our work, we performed BAR-Seq analyses on *in vitro* cultured samples of edited HSPCs, on whole blood samples collected at different times after transplant and on sorted human cell lineages (B cells, myeloid cells, T cells, HSPCs) from hematopoietic organs of reconstituted NSG mice (either primary or secondary recipients)[9]. When analyzing blood samples or sorted cell lineages, it should be noted that the number of HDR-edited cells within the harvested bulk population might sometimes be low due to poor human cell engraftment, limited lineage output or low biological material. The BAR-Seq wet procedure and bioinformatic analysis successfully extracted BARs from a bulk population comprising as few as 100 HDR-edited cells. However, such a low number of cells may call for caution when interpreting this data since rare clones with <1% abundance may be undetectable. Based on our findings on the frequency of repopulating cells[9], collection of >10,000 bulk cells from sorted cell lineages should be set as threshold to obtain robust results even with as low as 1% HDR editing in the human graft.

In any case, equivalent amounts of gDNA should be used among different samples for library preparation to avoid biasing BAR-Seq analysis. BAR sequences can be extracted by PCR amplification. To minimize sequencing errors in the BAR region, we suggest designing asymmetric amplicons with the forward primer (Read 1, R1) binding close to the BAR sequence. Amplicon length can vary based on the position of the BAR in the HDR template. We usually design amplicons spanning a region of 300-400 bp to minimize carryover of primer and primer dimers during amplicon purification. BAR-Seq amplicon preparation is based on two PCR rounds of maximum 15-20 cycles each. In the first round, the BAR-containing region is amplified with a couple of "PCR1" primers designed to bind the target site (**Supplementary Table 3**). Of note, PCR1 primers and amplification conditions must be optimized for each target site of interest, which may introduce variability in amplification efficiency or sensitivity across different loci. In the second round, R1/R2 primer complementary sequences, i5/i7 Illumina indexes and P5/P7 are added to the amplicon by nested PCR using "PCR2" primers listed in **Supplementary Table 3**. Single-round PCR using only "PCR2" primers may be considered as an option for BAR extraction, although lower amplification efficiency may occur when performing BAR-Seq on few edited cells. We multiplexed up to 49 independent samples in the same Illumina sequencing run (for MiSeq: Reagent Kit v3). During amplicon preparation and sequencing, we suggest including one sample in which no BAR should be retrieved by the BAR-Seq bioinformatic analysis. This additional control may be helpful to evaluate the background sequencing noise, the presence of cross-contaminations and the extent of index switching. Higher or lower level of multiplexing are possible depending on the desired sequencing depth for each sample. Sequencing read length may vary according to the position of the BAR within the amplicon. Although in our key reference paper we performed paired-end sequencing (in which we discarded the R2 file because only the R1 reads contained the amplicon), single-end Illumina MiSeq, NextSeq or HiSeq sequencing is sufficient since the bioinformatic pipeline works with only one fastq file for each sample. In fact, the only requirement is that the amplicons containing the BARs are fully contained in the sequencing reads provided to the pipeline. In case the full amplicon is too long to be fully contained in a single Illumina read, paired-end sequencing can be performed so that both reads will cover the BAR sequence. In this case we suggest performing a read merging using FLASh[47], which is a software specifically designed to merge pairs of reads when the original DNA fragments are shorter than twice the length of reads. The resulting longer reads can be then provided to the BAR-Seq pipeline.

Importantly, also sequencing depth strictly depends on the number of edited cells analyzed. In agreement with other protocols[2], we advise sequencing about 100 reads for each HDR-edited cell and we recommend avoiding massive over- or under-sequencing of the samples, which may increase

the background noise or may miss some clones, respectively. When applied to clonal tracking of gene edited HSPCs *in vivo*, from 50,000 to 500,000 reads for each sample are sufficient (considering the average HDR editing efficiency and the advised number of cells to collect). Replicates of sample library preparation and sequencing are typically not necessary but may be relevant when analyzing samples with low input gDNA in order to minimize sampling issues.

BAR-Seq bioinformatic analysis. The BAR-Seq bioinformatic pipeline is freely available at https://bitbucket.org/bereste/bar-seq. After download and installation of the required software and packages, the pipeline can be executed locally (**Figure 3**).

As in the majority of the NGS reads analyses, the first bioinformatic operations to be performed are quality check, quality filtering and trimming (if needed) of input sequences. As first goal, the BAR-Seq bioinformatic pipeline extracts BARs from sequencing reads by exploiting the amplicon structure, i.e. the known conserved sequences flanking the BAR (upstream and downstream). BAR-Seq uses TagDust2[48] to process the input reads and extract BAR sequences, since its Hidden Markov Model (HMM) for complex sequence structures, which include gaps and partial blocks, provides good flexibility in the detection of anchor sequences (provided as input) used to identify the BAR, even in presence of errors in the sequences around the BAR. The HMM-based approach implemented in TagDust2 is quite helpful to retrieve BARs from reads in which sequencing errors in the known amplicon sequences adjacent to the BAR are present. To balance this flexibility, we imposed an additional filter based on the barcode structure and its length (see **Design and production of the barcoded plasmid library** section) to discard all the extracted BARs not satisfying these constraints. Although by default our pipeline relies on TagDust2, any bioinformatic tool able to identify unknown substrings of variable length by exploiting the known adjacent ones, possibly containing mismatches or small insertions/deletions, can be used to accomplish this task. A notable example of such applications is the R package genBaRcode, which allows to digest sequencing reads with many different error-correction approaches and visualization routines[49]. Anyway, as a final result of this step, the preliminary set of BARs is extracted from the input sequences and their abundance is computed by counting the number of occurrences.

Since BARs extracted with TugDust2 could have different lengths due to its flexibility, a preliminary filter based on sequence length distribution estimates the most recurring BAR length value and keeps only those BARs having that length, while discarding those too long or too short. Moreover, the BAR-Seq pipeline offers the possibility to apply or not additional structural filters. In the former case, the user has to possibility to filter out: i) sequences having in at least one position a

406 nucleotide with frequency below 1%, assuming that a nucleotide with such a low nucleotide
407 abundance is artifactual, or ii) BARs not respecting a fixed structure (expressed in IUPAC code).
408 However, the number and the counts of this preliminary set of BARs could be influenced by
409 errors occurring during the sequencing process. Indeed, sequencing errors could produce low count
410 spurious BARs bearing sequence similarity with the much more abundant ones. To account for such
411 issue, we developed a graph-based method which identifies and then merges "ego-networks" for each
412 independent sample to obtain a refined set of BARs. Briefly, a graph is built from the pairwise
413 Levenshtein distance d between BARs, including only edges having d lower than a defined threshold
414 t ($d \leq 3$ in our experiments). While higher thresholds tend to produce spurious results and merge
415 unrelated sequences, the appropriate value t may depend on BAR length. An ego-network is created
416 for each node, i.e. a subgraph where the BAR of interest (the "focal" node) is connected to related
417 sequences (the "alters") having $d \leq t$ by definition. Each ego-network is then collapsed to the focal
418 node and the alters are removed from the graph. This process is performed iteratively on all BARs
419 ranked in descending abundance. Importantly, this approach does not build shared "ego-networks"
420 across samples, even when providing multiple samples at once. Of note, we avoided to use a
421 predefined list of known sequences ("whitelist") for the extraction of BARs because it would require
422 nearly exhaustive sequencing of the original plasmid/viral library and likely multiple rounds of library
423 sequencing to minimize the loss of real BARs and the inclusion of artifactual BARs in the whitelist,
424 respectively. Indeed, an incomplete whitelist may compromise the next step of the BAR-Seq analysis
425 by dropping out real barcodes (and therefore clones) and underestimating cell population complexity.
426 Optionally, further filtering can be applied to this refined set of BARs by discarding those
427 having read count lower than a selected threshold c ($c < 3$ in our experiments[9]). This filtering might
428 be useful to take into account background sequencing noise. To verify the need of this threshold and
429 estimate the appropriate value c, we strongly suggest including in the sequencing run a control
430 amplicon derived from cells edited with the conventional non-barcoded HDR template, where no
431 BAR should be retrieved. After filtering, the remaining BARs can be identified as the set of valid
432 BARs. To verify that this set of BARs is informative, we improved the approach previously described
433 to estimate the level of richness for each sample[9,50]. Briefly, the richness is defined as the ratio
434 between the number of observed BARs and the Chao1 abundance-based index, corrected to take into
435 account the minimum count threshold c. Optimal experiments should achieve high richness values
436 ($R>0.7$), while low richness levels ($R<0.3$) could indicate poor sequencing depth. In this case, the
437 amplicon preparation or sequencing process must be retried following our **Troubleshooting** guide.
438 Moreover, we provided additional metrics to better describe the sample complexity: Effective
439 Number of Species (ENS) and Equitability[28].

Since most studies are focused on highly abundant clones more robustly contributing to the cell population output, BAR-Seq gives to possibility to identify and select the (sub)set of "dominant" BARs. Valid sequences are ranked from the most to the least abundant and a saturation-based approach is applied. The dominant set is defined as the pool of BARs representing more than a chosen ratio of the total abundance, while the remaining BARs are considered as "rare". The user can select the value for the identification of dominant BARs based on their distribution. In our BAR-Seq analyses on repopulating edited cells in xenotransplanted hosts, we set this value to 90 or 95%[9].

When analyzing more samples, BAR sharing (i.e. the abundance of each valid sequence in all samples) can be computed. BAR sharing is reported in tabular form with valid BARs as rows and samples as columns, which gives the possibility to check for contaminations among samples or longitudinal BAR tracking within the same experimental condition.

To facilitate the analyses, we developed a user-friendly web-application available at http://www.bioinfotiget.it/barseq, which can perform the whole pipeline remotely upon the upload of the input sequencing files and the description of the amplicon structure by specifying the conserved sequences flanking the BAR (upstream and downstream) (**Figure 4**). In the home page of this online tool, the user can customize the aforementioned thresholds and values for background sequencing noise and dominant BARs identification. Further parameters can be customized by clicking on the "Advanced Options" button. It must be noticed that the application requires that all the pre-processing steps, like quality filtering or trimming of the input sequences, are performed prior to the upload. The pipeline runs on a dedicated server and results are stored for 48 hours. Results are available for visualization and download using the provided link. Moreover, an email notification is sent to the user when the analyses are finished. For each submitted sample, the result page presents a report of the sequencing run with the number of successfully or unsuccessfully extracted BARs and the statistics about each step of the BAR-Seq pipeline. Moreover, the set of valid BARs as well as the dominant ones are shown and can be downloaded as text files ("Download Full/Selected Barcode Results"). Plots showing BAR structures and saturation curves can be downloaded as images (.png format) by clicking on them. Additionally, when providing multiple samples, the sharing of valid and dominant BARs is computed and available in tabular form as well as heatmap ("Shared Results"). Sample diversity and richness can be automatically measured with three different indexes (ENS, Equitability and Chao1 abundance-based index) by clicking on the "Check Diversity/Richness" button in the "Results" page. Alternatively, the same operation can be done manually by downloading the text file from the "Results" page and uploading it in the "Check Diversity/Richness" tab.

Of note, when analyzing the original plasmid/viral library, the latter BAR-Seq web-application function allows to estimate the number of uniquely labeled cells that can be tracked with

a certain confidence level by using the given library (as shown in the resulting table). Two more tabs are available on the website: “Help” for more details concerning the use of the different web tools with related examples to test their functionality; “Contacts” for reporting any issue related to the use of the web application.

LIMITATIONS

The efficiency of HDR may limit BAR-Seq application particularly in slowly cycling or quiescent cells. HDR editing also requires extensive manipulation (DNA DSB and simultaneous delivery of the DNA template), which might cumulatively impact on cell survival and proliferation. These limitations, however, pertain to the biology of target cells and their suitability for templated editing and not to the tracking technology *per se*. Several strategies to enhance HDR editing have been proposed so far[13]. In our study we found that transient hyperactivation of the E2F pathway and simultaneous dampening of the editing-induced p53-mediated response increase the permissiveness to HDR in long-term repopulating HSPCs and improve the tolerability of the editing procedure[9,42].

Although the frequency of biallelic HDR targeting is generally low, a fraction of edited cell clones may carry two BARs if targeting somatic chromosomes. Such multiple BAR integrations might have minimal influence on data interpretation when interrogating cell fate and clonal composition[27]. Editing of sex-linked chromosomes in male cells allows more accurate quantification of the clonal composition.

BAR-Seq does not provide information on the dynamics and clonality of unedited or NHEJ-edited cells, which may be present in indifferent proportions in a cell population treated for HDR-editing. In our study, we combined BAR-Seq and CRISPResso2[45] pipelines to comprehensively analyze clonal behavior of HDR- and NHEJ- edited cells[9]. Finally, detection of rare quiescent or short-living cells that provide very limited cell output in transplantation experiments might be challenging with the BAR-Seq pipeline due to their low abundance close to sequencing background noise.

Low sequencing depth may result in an under-sampling of the overall population of cells (and, consequently, of BARs) in the sample, limiting the overall BAR-Seq procedure. In this conditions, detection of low abundant BARs might be challenging, especially when analyzing highly polyclonal populations.

MATERIALS

1.1. REAGENTS

Barcoded HDR plasmid cloning

- HPLC purified single-stranded oligodeoxynucleotide (ODN) containing the degenerated BAR sequence, the stuffer sequences and the restriction site as in **Fig. 2c** (Sigma-Aldrich, or another vendor); an example of the BAR ssODN is provided in **Supplementary Table 3**
- Non-barcoded plasmid template (custom). Synthesis of the non-barcoded HDR template and subcloning can be performed by gene synthesis services (GenScript; or another vendor). Plasmid backbones suitable for subcloning are any transfectable plasmid (e.g. pUC19), pAAV-MCS (Agilent; cat no. 240071) or pCCL-LV (available under Material Transfer Agreement with Ospedale San Raffaele/Telethon, Luigi Naldini's lab)
- Bsu36I restriction enzyme (New England Biolabs, cat. no. R0524S)
- SphI-HF® restriction enzyme (New England Biolabs, cat. no. R3182S)
- CutSmart® buffer (New England Biolabs, cat. no. B7204S; typically provided with restriction enzymes)
- SeaKem® GTG® Agarose (Lonza, cat. no. 50070)
- Gel Loading Dye, Purple (New England Biolabs, cat. no. B7024S)
- 1 kb DNA Ladder (New England Biolabs, cat. no. N3232S)
- Atlas ClearSight DNA Stain (Bioatlas, cat. no. BH40501)
- NucleoSpin® Gel and PCR Clean-up (Machery-Nagel, cat. no. 740609.250S)
- Easy-A High-Fidelity PCR Master Mix (Agilent Technologies, cat. no. 600640)
- Primers (Sigma-Aldrich or another vendor; see **Supplementary Table 3** for ordering details)
- MinElute PCR Purification Kit (Qiagen, cat. no. 28004)
- QIAquick PCR Purification Kit (Qiagen, cat. no. 28104)
- TAE (50x concentrate) pH8.0 (VWR, cat. no. 444125D)
- Antarctic Phosphates (New England Biolabs, cat. no. M0289L)
- T4 DNA Ligase (New England Biolabs, cat. no. M0202L)
- High Sensitivity D1000 ScreenTape (Agilent Technologies, cat. no. 5067-5585)
- High Sensitivity D1000 Reagents (Agilent Technologies, cat. no. 5067-5584)
- XL10-Gold Ultracompetent Cells (Agilent Technologies, cat. no. 200315)
- Nuclease-free Water for Molecular Biology (Sigma-Aldrich, W4502-1L)
- S.O.C. Medium (Invitrogen, cat. no. 15544034)
- LB Broth (ThermoFisher, cat. no. 10855001)
- Ampicillin (Sigma-Aldrich, cat. no. 10835242001) or another antibiotic accordingly to plasmid resistance.
- 150 mm LB Agar Plates with 100 µg/ml Ampicillin (VWR; cat. no. 100217-214)

- NucleoBond Xtra Maxi EF (Machery Nagel, cat. no. 740424.50)

***Ex-vivo* culture of human HSPCs**

- Iscove's DMEM, 1X (Corning, cat. no. 10-016-CV)
- FetalClone® II (HyClone™, cat. no. SH30066.03)
- DPBS without Ca^{2+} and Mg^{2+}, 1X (Corning, cat. no. 21-031-CV)
- ACK lysing buffer (Gibco, cat. no. A10492-01)
- StemSpan SFEM (StemCell Technologies, cat. no. 09650)
- Penicillin – Streptomycin (Lonza, cat. no. DE17-602E)
- L-Glutamine (Lonza, cat. no. 17-605E)
- Recombinant human SCF (PeproTech, cat. no. 300-07)
- Recombinant human Flt3-L (PeproTech, cat. no. 300-19)
- Recombinant human TPO (PeproTech, cat. no. 300-18)
- Recombinant human IL-6 (PeproTech, cat. no. 200-06)
- 16,16-dimethyl prostaglandin E2 - PGE_2 (Cayman, cat. no. 14750).
- StemRegenin 1 - SR1 (Biovision, cat. no. 1967)
- UM171 (StemCell Technologies, cat. no. 72914)

Gene editing of human HSPCs

Synthetic chemically modified guide RNAs (2 part):

- Alt-R® CRISPR-Cas9 tracrRNA, 100 nmol (IDT, cat. no. 1072534)
- Alt-R CRISPR-Cas9 crRNA, 10 nmol (IDT, custom)

CRITICAL STEP – Alternatively, single gRNAs (Synthego, custom; or another vendor) can be used.

- Alt-R® Cas9 Electroporation Enhancer, 10 nmol (IDT, cat. no. 1075916)
- sNLS-SpCas9-sNLS Nuclease (Aldevron, cat. no. 9212-5MG)
- P3 Primary Cell 4D-NucleofectorTM X Kit S (Lonza, cat. no. V4XP-3032)
- DPBS, 1x without calcium and magnesium (Corning, cat. no. 21-031-CV)

Xenotransplantation of edited HSPCs

- DPBS, 1x without calcium and magnesium (Corning, cat. no. 21-031-CV)

***In-vivo* phenotyping of human edited lineages (GFP compatible)**

- 0.5 M EDTA, pH 8.0 (Invitrogen, cat. no. 15575-038) > dilution 1:3 in PBS
- SPHERO™ Rainbow Calibration Particles (Spherotech, cat. no. RCP-30-5A)

- BD FACS™ Accudrop beads (BD; cat. no. 345249)

Antibodies for peripheral blood and spleen phenotyping

- Anti-human CD45-PB, Clone HI30 (BioLegend, cat. no. 304029, RRID: AB_2174123) https://scicrunch.org/resolver/RRID:AB_2174123
- Anti-human CD19-PE, Clone SJ25C1 (BD Biosciences, cat. no. 345789, RRID: AB_2868815) https://scicrunch.org/resolver/RRID:AB_2868815
- Anti-human CD13-APC, Clone WM15 (BD Biosciences, cat. no. 557454, RRID: AB_398624) https://scicrunch.org/resolver/RRID:AB_398624
- Anti-human CD3-PE-Cy7, Clone HIT3a (BioLegend, cat. no. 300316, RRID: AB_314052) https://scicrunch.org/resolver/RRID:AB_314052
- Anti-human CD8-APC-H7, Clone Sk1 (BD Biosciences, cat. no. 641400, RRID: AB_1645736) https://scicrunch.org/resolver/RRID:AB_1645736
- 7-AAD Viability Staining Solution (BioLegend, cat. no. 420404)

Antibodies for bone marrow phenotyping

- Anti-human CD45-APC-H7, Clone 2D1 (BD Biosciences, cat. no. 348815, RRID: AB_2868859) https://scicrunch.org/resolver/RRID:AB_2868859
- Anti-human CD34-PE-Cy7, Clone 8G12 (BD Biosciences, cat. no. 348811, RRID: AB_2868855) https://scicrunch.org/resolver/RRID:AB_2868855
- Anti-human CD19-PE, Clone SJ25C1 (BD Biosciences, cat. no. 345789, RRID: AB_2868815) https://scicrunch.org/resolver/RRID:AB_2868815
- Anti-human CD33-VioBlue, Clone AC104.3E3 (Miltenyi Biotec, cat. no. 130-099-485, RRID: AB_2660351) https://scicrunch.org/resolver/RRID:AB_2660351
- Anti-human CD3-APC, Clone UCHT1 (BD Biosciences, cat. no. 555335, RRID: AB_398591) https://scicrunch.org/resolver/RRID:AB_398591
- 7-AAD Viability Staining Solution (BioLegend, cat. no. 420404)

BAR-Seq library preparation and quantification

- QIAamp DNA Micro Kit (Qiagen, cat. no. 56304)
- Primers (Sigma-Aldrich, Metabion or another vendor; see **Supplementary Table 3** for ordering details)
- GoTaq® G2 DNA Polymerase (Promega, cat. no. M7841) or Pfu DNA Polymerase (Promega, cat. no. M7741)

- Nuclease-free Water for Molecular Biology (Sigma-Aldrich, W4502-1L)
- dNTP Mix (Promega, cat. no. U1515)
- AmpPure XP (Beckman Coulter, cat. no. A63881)
- High Sensitivity D1000 ScreenTape (Agilent Technologies, cat. no. 5067-5585)
- High Sensitivity D1000 Reagents (Agilent Technologies, cat. no. 5067-5584)

Illumina sequencing

- PhiX control V3 (Illumina, cat. no. FC-110-3001)
- MiSeq™ Reagent Kit v3 (150-cycle) (Illumina, cat. no. MS-102-3001) or NextSeq 500/550 High Output Kit v2.5 (150 Cycles) (Illumina, cat. no. 20024907)

1.2. BIOLOGICAL MATERIAL

- Frozen CB-derived $CD34^+$ human HSPCs (Lonza, cat. no. 2C-101; or another vendor).
  ! CAUTION – *The use of human material must be approved by local institution and authorities.*
- NOD.Cg-Prkdc^{SCID} Il2rg$^{tm1Wjl/SzJ}$ (NSG) female mice (Charles River, cat. no. NSG614)
  ! CAUTION – *The use of animal models must be approved by local institution and authorities in agreement with law.*

1.3. EQUIPMENT

1.3.1. Experimental part

- Pipet-Lite LTS Pipette L-1000XLS+ (Rainin, cat. no. 17014382)
- Pipet-Lite LTS Pipette L-200XLS+ (Rainin, cat. no. 17014391)
- Pipet-Lite LTS Pipette L-20XLS+ (Rainin, cat. no. 17014392)
- Pipet-Lite LTS Pipette L-10XLS+ (Rainin, cat. no. 17014388)
- Pipet-Lite XLS+ Multichannel Pipette 8ch/10 µL (Rainin, cat. no. 17013802)
- Pipet-Lite XLS+ Multichannel Pipette 8ch/20 µL (Rainin, cat. no. 17013803)
- Pipet-Lite XLS+ Multichannel Pipette 8ch/200 µL (Rainin, cat. no. 17013805)
- Pipette Tips RT UNV 10µL FGL 960A/10 (Rainin, cat. no. 30389176)
- Pipette Tips RT LTS 20µL FL 960A/10 (Rainin, cat. no. 30389226)
- Pipette Tips RT LTS 200µL FLW 960A/10 (Rainin, cat. no. 30389241)
- ART 10 Pipette Tips (Carlo Erba Reagents, cat. no. FL9214000)
- ART 20P Pipette Tips (Carlo Erba Reagents, cat. no. FL92149P0)
- ART 200 Pipette Tips (Carlo Erba Reagents, cat. no. FL9206900)

- ART 1000 Pipette Tips (Carlo Erba Reagents, cat. no. FL92079E0)
- SWIFTPET PRO (HTL, cat. no. 0390)
- Falcon® 2 mL Serological Pipet (Corning, cat. no. 357507)
- Falcon® 5 mL Serological Pipet (Corning, cat. no. 357543)
- Falcon® 10 mL Serological Pipet (Corning, cat. no. 357551)
- Falcon® 25 mL Serological Pipet (Corning, cat. no. 357525)
- Falcon® 96-well Clear Microplate (Corning, cat. no. 353072)
- Costar® 48-well Clear TC-treated Multiple Well Plates (Corning, cat. no. 3548)
- Falcon® 24-well Clear Flat Bottom TC-treated Polystyrene Multiwell Cell Culture Plate (Corning, cat. no. 353047)
- Falcon® 6-well Clear Multiwell Plate (Corning, cat. no. 353046)
- Nunclon™ Delta Surface 12-well Plate (Thermo Fisher Scientific, cat. no. 150628)
- 1000 µL Filtered tips for pipettes (Thermo Fisher Scientific, cat. no. FL92079E0)
- 200 µL Filtered tips for pipettes (Thermo Fisher Scientific, cat. no. FL9206900)
- 20 µL Filtered tips for pipettes (Thermo Fisher Scientific, cat. no. FL92149P0)
- 10 µL Filtered tips for pipettes (Thermo Fisher Scientific, cat. no. FL9214000)
- Thermowell® Tube Strip (Corning, cat. no. 6542)
- Eppendorf Safe-Lock tubes, 0.5 mL (Eppendorf, cat. no. 0030121023)
- Eppendorf Safe-Lock tubes, 1.5 mL (Eppendorf, cat. no. 0030120086)
- DNA LoBind Tubes, 1.5 mL (Eppendorf, cat. no. 0030108051)
- DNA LoBind Tubes, 2.0 mL (Eppendorf, cat. no. 0030108078)
- Eppendorf twin.tec® PCR plate 96 LoBind (Eppendorf, cat. no. 0030129504)
- Falcon™ 50mL Conical Centrifuge Tubes (Fisher Scientific, cat. no. 14-432-22)
- Falcon™ 15mL Conical Centrifuge Tubes (Fisher Scientific, cat. no. 11507411)
- Falcon™ Round-Bottom Polystyrene Test Tubes with Cell Strainer Snap Cap (Fisher Scientific, cat. no. 08-771-23)
- L-Shape Cell Spreader (VWR; cat. no. 76207-748)
- IKA® Vortex Genius 3 (IKA, cat. no. 00033400009)
- IKA® MS3 basic (IKA, cat. no. 0003617000)
- ThermoMixer® C (Eppendorf, cat. no. 5382000015)
- Centrifuge 5430 (Eppendorf, cat. no. 5427000415)
- Centrifuge 5810R (Eppendorf, cat. no. 5811000620)
- Disposable scalpel, stainless steel blade no. 11 (VWR, cat. no. 1110811)
- T100™ Thermal Cycler (Bio-Rad, cat. no. 1861096)

- Heracell™ 150i CO_2 Incubator (Thermo Fisher Scientific, cat. no. 50116048)
- New Brunswick™ Innova® 42 (Eppendorf, cat. no. M1335-0012)
- Nanodrop™ 8000 Spectrophotometer (Thermo Fisher Scientific, cat. no. ND-8000-GL)
- Eppendorf Biophotometer® D30 (Eppendorf, cat. no. 6133000001)
- PowerPac™ Basic Power Supply (Bio-Rad, cat. no. 1645050)
- Sub-Cell® GT Gel Tray (Bio-Rad, cat. no. 1704416)
- Sub-Cell® GT Comb 10-well (Bio-Rad, cat. no. 170444)
- Wide Mini-Sub Cell® GT Horizontal Electrophoresis System (Bio-Rad, cat. no. 1704468)
- Analytical Balance ME54T/00 (Mettler Toledo, cat. no. 30216540)
- Molecular Imager® Gel Doc™ XR System (Bio-Rad, cat. no. 1708195EDU)
- QUICKSPIN2 Mini-Centrifuge (Cleaver Scientific, cat. no. 5055323269776)
- Agilent 4200 TapeStation System (Agilent Technologies, cat. no. G2991AA)
- Loading Tips (Agilent Technologies, cat. no. 5067-5599)
- Pierceable Foil Heat Seal for PCR applications (Bio-Rad, cat. no. 1814040)
- 4D-Nucleofector™ Core Unit (Lonza, cat. no. AAF-1002B)
- 4D-Nucleofector™ Core Unit (Lonza, cat. no. AAF-1002X)
- Falcon® 40 µM Cell Strainer (Corning, cat. no. 352340)
- Falcon® 5 mL Round Bottom Polystyrene Test Tube with Cell Strainer Snap Cap (Corning, cat. no. 352235)
- BD FACSAria Fusion (BD, cat. no. special order)
- BD FACSCanto II (BD, cat. no. special order)
- FCS Express 6 Flow (De Novo Software, https://denovosoftware.com/purchasing)
- DynaMag™ 96 Side Skirted Magnet (Thermo Fisher Scientific, cat. no. 12027)
- TQ-Prep™ Workstation (Beckman Coulter, cat. no. 6605429)
- IMMUNOPREP Reagent System (Beckman Coulter, cat. no. 7546999)
- TC20™ Automated Cell Counter (Bio-Rad, cat. no. 1450102)
- Cell Counting Kit (Bio-Rad, cat. no. 1450003)
- RS-2000 X-ray Irradiator (Rad Source, cat. no. special order)
- PX1 PCR Plate Sealer (Bio-Rad, cat. no. 1814000)
- Microvette® 500 K3E (Sarstedt, cat. no. 20.1341)
- BD Micro-Fine+ 0.5 ml U-100 insulin syringe (BD, cat. no. 324892)

1.3.2. Computational part

- The pre-processing of sequences can be performed on a computer running the Linux terminal (also called command line interface or shell), with installed common utilities for the read quality control and cleaning. A common approach is to use FastQC https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ for read quality control and Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) for read filtering and trimming. Details about the installation and requirements of these applications on different Operating Systems are available in their reference manuals. The required computational resources depend on the number of sequencing reads, even if a commodity laptop is usually enough to perform all the computation.
- The BAR-Seq pipeline can be run locally on a Linux computer with Python3 installed and the following pre-requisite packages: numpy, editdistance, network, math, pandas, matplotlib, logomaker (as reported in the github repository https://bitbucket.org/bereste/bar-seq). Moreover, the pipeline requires the TagDust2 software installed (http://tagdust.sourceforge.net). Details about the installation and requirements of these applications on different Operating Systems are available in their reference manuals. The required computational resources depend on the number of sequencing reads, even if a commodity laptop is usually enough to perform all the computation.
- Alternatively, the BAR-Seq pipeline can be run online using a common laptop with internet access (www.bioinfotiget/barseq) and no other specific software installed. A detailed description of the web application usage is present on the website.

1.4. REAGENT SETUP

HSPC medium preparation

! CAUTION – *This reagent preparation must be performed in sterile hood.*

For *ex-vivo* culture of CB CD34$^+$ HSPCs, serum-free StemSpan™ SFEM medium should be supplemented with 100 IU/mL penicillin, 100 µg/mL streptomycin, 2% (vol/vol) glutamine, 100 ng/mL hSCF, 100 ng/mL hFlt3-L, 20 ng/mL hTPO, 20 ng/mL hIL-6, 1 µM SR1, and 50 nM UM171. Medium can be stored at 4°C for maximum 1 week.

Cas9 RNP preparation

! CAUTION – *This reagent preparation must be performed in sterile hood.*

To obtain the required amount of gRNA, mix custom Alt-R® CRISPR-Cas9 crRNA and tracrRNA with a 1:1 ratio, incubate at 95 °C for 5 min and keep cooling at room temperature for 10 min. If single gRNA (purchased by Synthego, or other vendor) is used, there is no need for the annealing

step. To prepare 25 pmol of RNP complex, add in a new sterile Eppendorf tube 1 µL DPBS, 0.41 µL of Cas9 protein and then gRNA to a predefined gRNA:Cas9 molar ratio (typically ≥1.5). Incubate at room temperature for 15 min to allow complexing. Add 1 µL of Alt-R® Cas9 Electroporation Enhancer (0.1 nmol). If single gRNA is used, there is no need to add the electroporation enhancer. We advise to prepare fresh Cas9 RNP complex.

P3 mix supplementation

! CAUTION – *This reagent preparation must be performed in sterile hood.*

Add one vial of Supplement 1 Solution (Lonza) to one vial of P3 Primary Cell Nucleofector™ Solution (Lonza). Briefly mix by vortexing. The supplemented solution can be stored at 4°C for maximum 3 months.

Antibody cocktail for *in vivo* peripheral blood and spleen phenotype

Target	Fluorochrome	Working Dilution	Amount x 100µl (each reaction)
CD45$^+$	PB	(1:50)	2 µL
CD19$^+$	PE	(1:50)	2 µL
CD13$^+$	APC	(1:50)	2 µL
CD3$^+$	PE-Cy7	(1:50)	2 µL
CD8$^+$	APC-H7	(1:50)	2 µL

We advise to prepare fresh antibody cocktail.

Antibody cocktail for *in vivo* bone marrow phenotype

Target	Fluorochrome	Working Dilution	Amount x 100µl (each reaction)
CD45$^+$	APC-H7	(1:50)	2 µL
CD34$^+$	PE-Cy7	(1:50)	2 µL
CD19$^+$	PE	(1:50)	2 µL
CD33$^+$	PB	(1:50)	2 µL
CD3$^+$	APC	(1:50)	2 µL

We advise to prepare fresh antibody cocktail.

1.5. EQUIPMENT SETUP

BD FACSAria Fusion

For cell sorting, BD FACSAria Fusion should be equipped with four lasers: blue (488 nm), yellow/green (561 nm), red (640 nm) and violet (405 nm). We advise to use an 85 µm nozzle and set sheath fluid pressure at 45 psi. We recommend a highly pure sorting modality (4-way purity sorting). The drop delay should be determined by BD FACS Accudrop beads prior to sorting.

PROCEDURE

Preparation of the BAR-Seq HDR template library TIMING **– 4 days**

1 Synthetize the complementary strand and amplify the barcoded ssODN using appropriate primers (see **Supplementary Table 3**). Set up an Easy-A High-Fidelity PCR master mix for 5 reactions (+1 without barcoded ssODN template as negative control):

Component	Amount (each reaction)	No. of reactions	Total amount
Barcoded ssODN	50 pmol	5	250 pmol
10X-Easy-A buffer	5 µL		25 µL
dNTPs (10 mM)	1.2 µL		6 µL
Library Fw (100 µM)	0.5 µL		2.5 µL
Library Rv (100 µM)	0.5 µL		2.5 µL
Easy-A enzyme	0.5 µL		2.5 µL
Molecular-grade H_2O	Up to 50 µL		Up to 250 µL

Briefly mix and spin, and then load on a thermocycler machine:

Step	Number of cycles	Temperature	Duration
1	1	95°C	2 min
2	10	95°C	40 sec
		60°C	30 sec
		72°C	30 sec
3	1	72°C	5 min
4	1	4°C	-

2 Pool the five reactions and purify them with the MinElute PCR Purification Kit, accordingly to manufacturer's instructions (elution volume = 10 ul).

3 Quantify the purified product with NanoDrop 8000.

4 Set up digestion reactions of the purified product with restriction enzymes Bsu36I and SphI. Keep 2 µl of the purified undigested product as control.

Component	Amount
DNA (from step 2)	8 µL
10X CutSmart buffer	5 µL
Bsu36I	2 µL
SphI	2 µL
Molecular-grade H_2O	23 µL

5 Incubate at 37°C for 60 min.

6 Purify the digestion with the MinElute PCR Purification Kit, accordingly to manufacturer's instructions (elution volume = 11 ul).

7 Equilibrate at room temperature the High Sensitivity D1000 Reagents and ScreenTape for 20 min.

8 Vortex High Sensitivity D1000 Reagents for 30 sec.

9 Separately prepare 1:10 and 1:100 dilution of the digested and undigested products; mix by briefly vortexing.

10 To verify successful digestion, load on Agilent 4200 TapeStation the dilutions of the digested and undigested products following manufacturer's instructions. In our experience, partial digestion may occur without major impact on the next steps.

? TROUBLESHOOTING

11 Use a NanoDrop spectrophotometer to quantify insert template concentration.

12 Set up digestion reactions of the non-barcoded plasmid backbone with restriction enzymes Bsu36I and SphI.

Component	Amount
Non-barcoded plasmid	40 µg
10X CutSmart buffer	10 µL

Bsu36I	5 µL
SphI	5 µL
Molecular-grade H_2O	Up to 100 µL

13 Purify the digested plasmid with QIAquick PCR Purification Kit, accordingly to manufacturer's instructions.

14 Set up 5' dephosphorylation of the digested plasmid with Antarctic Phosphatase.

Component	Amount
Digested plasmid (from step 12)	-
10X Antarctic P. buffer	10 µL
Antarctic Phosphatase	5 µL
Molecular-grade H_2O	Up to 100 µL

15 Incubate at 37°C for 60 min.

16 Prepare 1% agarose gel for plasmid purification upon electrophoresis. Premix the dephosphorylated and digested plasmid from step 13 with the loading dye and load it on the gel. At least 1 µg of undigested plasmid control and of 1 kb DNA ladder should also be loaded in separate wells to check for plasmid size and effective digestion.

17 Read the agarose gel on a GelDoc imager and isolate with a scalpel the portion of agarose gel embedding the DNA band of the dephosphorylated and digested plasmid.

! CAUTION – *Avoid direct body exposure to UV light. Always wear UV-light protective equipment.*

? TROUBLESHOOTING

18 Purify the DNA band with the NucleoSpin® Gel and PCR Clean-up, following manufacturer's instructions (elution volume = 30 µL). Use a NanoDrop spectrophotometer to quantify plasmid concentration.

PAUSE STEP – Store at -20°C for long-term storage of the plasmid backbone and the insert template.

19 To assess the plasmid:insert molar ratio allowing the higher ligation efficiency, set up three different ligation reactions.

Component	Amount

Purified plasmid (from step 17)	100 ng
Purified insert (from step 12)	1x, 5x, 10x plasmid moles equivalent
2X T4 ligation buffer	10 µL
T4 DNA Ligase	1 µL
Molecular-grade H_2O	Up to 20 µL

An additional ligation reaction without the purified insert should be also set up as control.

Component	Amount
Purified plasmid (from step 17)	100 ng
2X T4 ligation buffer	10 µL
T4 DNA Ligase	1 µL
Molecular-grade H_2O	Up to 20 µL

20 Gently mix by pipetting, briefly spin and incubate the reactions at 25°C for at least 60 min.

21 Pre-warm at 30°C SOC and LB medium, and 150 mm LB agar plates supplemented with the appropriate antibiotic.

22 Chill on ice two vials of XL10-Gold Ultracompetent Cells for 5-10 min. Split each vial in two separate pre-chilled Eppendorf tubes (4 tubes total).

23 Add 2 µL of β-mercaptoethanol to each tube containing ultracompetent cells and gently swirl the tubes.

CRITICAL STEP – *Avoid vortexing XL10-Gold ultracompetent cells.*

24 Incubate on ice for 10 min.

25 Add 5 µL of the ligation products to each ultracompetent cells tube and gently swirl.

CRITICAL STEP – *Avoid vortexing XL10-Gold ultracompetent cells.*

26 Incubate on ice for 30 min.

27 Heat-pulse the tubes at 42°C for 30 sec and then incubate on ice for 2 min.

CRITICAL STEP – *The duration of the heat-pulse is critical to achieve optimal transformation efficiency. We recommend not to overextend or shorten this time.*

28 Add 450 µL of prewarmed SOC medium to each tube and incubate them for 60 min at 37°C with shaking at 300 rpm.

29 For each tube containing the bacteria outgrowth from step 27, prepare 1:1, 1:10, 1:50 dilutions in pre-warmed LB medium (final volume = 300 µL) and plate them in pre-warmed 150 mm LB agar plates.

30 Incubate the plates at 30°C overnight.

31 Determine/estimate the total number of colonies/plate. This step allows to identify: i) the plasmid:insert molar ratio that ensures the higher transformation efficiency; ii) the dilution of bacteria outgrowth avoiding colonies overcrowding in the plate (DF); iii) the maximum estimated theoretical complexity of the barcoded library when transforming the whole product of one ligation reaction ("no. of colonies/plate" x "DF" x 4), and therefore the number of ligation reactions ("M") required to reach the target library complexity; iv) the number of XL10-Gold vials ("V") and 150 mm LB agar plates ("P") needed for the library production.

CRITICAL STEP – *We advise to count the number of colonies in a subarea of the plate and then multiply this number for the area:subarea ratio. The number of colonies may be uncountable at 1:1 and 1:10 dilutions. Furthermore, the estimated number of colonies in the control condition should be at least 10^4-fold lower than in the other plates to minimize the chance of having clones carrying the non-barcoded template.*

? TROUBLESHOOTING

32 Having identified the optimal plasmid:insert molar ratio for ligation, set up "M" ligation reactions by scaling up calculations from step 18.

Component	Amount
Purified plasmid (from step 17)	100 ng x M
Purified insert (from step 12)	Determined in step 30 x M
2X T4 ligation buffer	10 µL x M
T4 DNA Ligase	1 µL x M
Molecular-grade H_2O	Up to 20 µL x M

33 Gently mix by pipetting, briefly spin and incubate the reactions at 25°C for at least 60 min.

34 Pre-warm at 30°C SOC and LB medium and 150 mm LB agar plates supplemented with the appropriate antibiotic.

35 Chill on ice a total number of "V" vials (as determined in step 30) of XL10-Gold ultracompetent cells for 5-10 min.

36 Add 4 µL of β-mercaptoethanol to each tube containing ultracompetent cells and gently swirl the tubes.

CRITICAL STEP – *Avoid vortexing XL10-Gold ultracompetent cells.*

37 Incubate on ice for 10 min.

38 Add 10 µL of the ligation products to each ultracompetent cells tube and gently swirl.

CRITICAL STEP – *Avoid vortexing XL10-Gold ultracompetent cells.*

39 Incubate on ice for 30 min.

40 Heat-pulse the tubes at 42°C for 30 sec and then incubate on ice for 2 min.

CRITICAL STEP – *The duration of the heat-pulse is critical to achieve optimal transformation efficiency. We recommend not to overextend or shorten this time.*

41 Add 900 µL of pre-warmed SOC medium to each tube and incubate them for 60 min at 37°C with shaking at 300 rpm.

42 Pool the bacteria outgrowths from step 42 in 15- or 50-mL Falcon tubes. Add pre-warmed LB medium to reach the predetermined "DF" dilution and plate in "P" 150 mm LB agar plates (as determined in step 30).

43 Incubate the plates at 30°C overnight and verify that the number of colonies is in agreement with the calculations from step 31.

44 Add 3 mL of LB medium (supplemented with the appropriate antibiotic) on the top of each LB agar plate. Scrape bacteria colonies by using bacteria spreaders. Collect, pool and thoroughly mix by pipetting the LB medium derived from each plate.

45 Inoculate 20-50 mL of the scraped product in 500 mL of LB medium and incubate at 30°C for additional 6 hours.

46 Perform plasmid extraction and purification from the bacteria outgrowth with the NucleoBond Xtra Maxi EF following manufacturer's instructions. Use a NanoDrop spectrophotometer to quantify plasmid concentration.

47 (Optional) Set up enzymatic digestion of 1 µg of the BAR-Seq plasmid library to check for plasmid integrity by gel electrophoresis. This step is strongly recommended in case the plasmid contains repeated sequences, such as AAV ITRs.

PAUSE STEP – *Store at -20°C for long-term storage.*

CRITICAL STEP – *Assess the diversity of the BAR-Seq plasmid library by deep-sequencing the BAR region (**Box 1**). Proceed to the next steps only if the library complexity and diversity are sufficient to univocally tag cell clones in the target population of interest (as discussed in the "Experimental design").*

? TROUBLESHOOTING

48 (Optional) The BAR-Seq plasmid library is ready to use in gene editing experiments. Alternatively, the BAR-Seq plasmid library may serve as transfer plasmid for BAR-Seq viral library production.

? TROUBLESHOOTING

CRITICAL STEP – *Assess the diversity of the BAR-Seq viral library before moving on to gene editing experiments by following steps provided in* ***Box 2****.*

Thawing of CB CD34$^+$ HSPCs TIMING – day '0' of the editing procedure– **30 min**

! CAUTION – *Cell culture and gene editing procedure must be performed in sterile hood.*

CRITICAL STEP – *CB CD34$^+$ HSPCs are purchased frozen from Lonza and contain at least 1x10^6 total cells. HSPCs must be conserved in liquid nitrogen.*

49 Determine the number of culture-initiating CB HSPCs required for the experiment.

CRITICAL STEP – *Transplantation of higher number of culture-initiating HSPCs in sub-lethally irradiated NSG mice results in the saturation of the hematopoietic niche. Therefore, the number of transplanted culture-initiating HSPCs is critical to potentially uncover differences in the clonality or repopulation capacity of edited cells. In our experience, below 1.5 x 10^5 and above 3 x 10^5 culture-initiating CB HSPCs/mouse are limiting or saturating ranges for cell doses, respectively.*

CRITICAL STEP – *If more than one vial is required, we strongly suggest to pool cells from different donors to reduce inherent variability.*

50 Pre-warm the supplemented StemSpan™ medium (see Reagent setup) at 37°C.

51 Thaw CB CD34$^+$ cells by immerging the vial in water bath at 37°C water for 5 min.

52 Transfer the cells solution in a 50-mL sterile Falcon tube, add "drop-by-drop" 10 volumes of DMEM medium and pellet them using the centrifuge 5810R (1100 rpm, room temperature, 10 min).

53 Carefully aspirate and discard the supernatant. Quickly and gently resuspend cell pellet with the pre-warmed StemSpan™ medium. Seed cells at the concentration 5 x 10^5 cells/mL.

CRITICAL STEP – *This cell concentration favorizes cell cycling during pre-stimulation while promoting cell maintenance.*

54 Add dmPGE$_2$ to the culture medium (final concentration = 10 µM) and mix well by pipetting.

CRITICAL STEP – *dmPGE$_2$ protects CB HSPCs from thawing toxicity and preserves their stemness properties*[18]*. dmPGE$_2$ should not be further supplemented in the medium after the editing procedure.*

55 Incubate cells for three days at 37°C in a 5% CO_2 and 20% O_2 humidified atmosphere.

CRITICAL STEP – *In our experience, three-days expansion of HSPCs before gene editing procedure allows to maximize HDR efficiency in the long-term repopulating HSPC compartment ([18] and unpublished data).*

Gene editing procedure of cultured HSPCs TIMING – day '+3' of the editing procedure – **1-2 hours**

56 Pre-warm the supplemented StemSpan™ medium at 37°C (see Reagent setup).

57 Count the number of viable cultured HSPCs, collect 1-5x10^5 cells in a 1.5-mL sterile Eppendorf tube, add 10 volumes of DPBS and pellet them using the centrifuge 5430 (2250 rpm, room temperature, 10 min).

58 Carefully aspirate and discard the supernatant. Resuspend cell pellet with P3 Primary solution mix (see Reagent setup) and add the RNP complex to reach a final volume of 20 µL/sample.
CRITICAL STEP –*The electroporation mixture may be supplemented with mRNA(s) to overexpress proteins of interest, as previously described for editing enhancers[9,42]. For optimal electroporation efficiency and lower cytotoxicity, the volume of the P3 primary solution mix should not be lower than 70% of the total electroporation volume. The final electroporation volume may be increased up to 25 µL/sample, if required. Be aware that increasing the electroporation volume above 25 µL/sample may cause electroporation failure.*

59 Transfer the cell solution to 4D-Nucleofector™ strip for each condition/mouse and proceed immediately to electroporation using the pre-recorded manufacturer's human $CD34^+$ **EO-100** program.

60 Wait 1 minute and then add 180 µL of pre-warmed StemSpan™ medium. Transfer cells to an adequate well-plate to reach the final concentration of 1 x 10^6 cells/mL.
CRITICAL STEP – *Cells belonging to the same experimental condition can be pooled.*

61 Incubate cells for **15 min** at 37°C in a 5% CO_2 and 20% O_2 humidified atmosphere.

62 Transduce electroporated cells with the BAR-Seq AAV6 library at multiplicity of infection (MOI) = 20.000 vg/cell and gently mix by pipetting.
CRITICAL STEP – *Keep the virus stock on ice. Do not thaw more than 3 times the same AAV aliquot to assure virus stability and reproducibility in results.*

63 Incubate cells for **24 hours** at 37°C in a 5% CO_2 and 20% O_2 humidified atmosphere.
CRITICAL STEP – In our experience, 24 hours is the minimum recovery time to avoid loss of cell engraftment after transplantation in NSG mice.

Transplantation of edited HSPCs in immunodeficient NSG mice TIMING – day '+4' of the editing procedure – **2 hours**

64 Irradiate NSG female mice at 190 rad at least 4 hours before cells transplantation.

CRITICAL STEP – We advise considering at least 5 mice per experimental group and ideally the experiment should be repeated twice to allow reliable statistical analysis. Mice should be randomly distributed across the experimental groups.

65 Count the number of viable edited HSPCs, collect all cells in a new 15-mL sterile Falcon tube, add 10 volumes of DPBS and pellet them using the centrifuge 5810R (1100 rpm, room temperature, 10 min).

66 Carefully aspirate and discard the supernatant. Resuspend cell pellet with DPBS to a final volume of 200 µL x no. mice to be transplanted/group + 50 µL excess.

67 Aliquot 200 µL in individual sterile 1.5 mL Eppendorf tubes (one for each mouse) and keep them on ice until transplantation.

68 Immediately proceed to transplant irradiated NSG mice. Perform intravenous tail-vein injection with the previously prepared 200 µL of DPBS containing edited HSPCs by using a sterile 0.5 mL insulin syringe with a 29-gauge x 12,7 mm needle.

CRITICAL STEP – Do not keep HSPCs on ice for too long to avoid cytotoxicity. Mouse tail may be pre-warmed with red-light lamp to ease injection. In case of potentially uncomplete or failed injection, the mouse should be univocally marked and must be excluded from the experiment in case engraftment failure is confirmed by flow cytometry analyses.

! CAUTION – Carefully handle syringes to avoid punctures.

69 Collect the remaining 50 µL of DPBS with edited HSPCs and add supplemented StemSpan™ medium to reach a final concentration of $5x10^5$ cells/mL. If the final volume is < 500 µL, pellet cells using the centrifuge 5430 (2250 rpm, room temperature, 10 min) before adding the medium and carefully remove the supernatant.

70 Culture HSPCs at 37°C in a 5% CO_2 and 20% O_2 humidified atmosphere for additional seven days to proceed with *in vitro* clonal analysis.

CRITICAL STEP – In vitro clonal analysis can be performed even after seven days of culture. We recommend caution when shortening this time of culture since barcoded template dilution may be uncomplete and thus PCR jumping may occur between on-target integrated barcodes and episomal barcoded AAV. In vitro clonal analysis may also be performed on sorted cell subpopulations.

71 (Optional) Further characterization of edited HSPCs *in vitro* may be relevant for the interpretation of clonality data, such as *in vitro* clonogenic capacity (**Box 3**), cell phenotype (**Box 4**), HDR/NHEJ editing efficiency (**Boxes 5 and 6**).

? TROUBLESHOOTING

Sample preparation for BAR-Seq analysis on cultured HSPCs TIMING – day '+10' of the editing procedure – **2 hours**

72 Collect >200,000 cultured HSPCs for each experimental condition, add 10 volumes of DPBS and pellet them using the centrifuge 5430 (2250 rpm, room temperature, 10 min).

73 Aspirate and discard the supernatant.

PAUSE STEP – *The cell pellet may be frozen at -80°C for long-term storage.*

74 Extract gDNA using the QIAamp DNA Micro Kit, accordingly to manufacturer's instructions. Use a NanoDrop spectrophotometer to quantify DNA concentration.

PAUSE STEP – *gDNA samples are now ready for BAR-Seq clonal analysis following steps from 104 to the end. gDNA can be frozen at +4 or -20°C for short- or long-term storage, respectively.*

Phenotypic analysis of peripheral blood samples from transplanted NSG mice TIMING – **3 hours / time point**

75 Every 3 weeks after the time of transplant, perform blood collection from transplanted NSG mice. For each mouse, prepare one 1.5-mL sterile Eppendorf tube with 20 µL of EDTA.

CRITICAL STEP – The size of the human cell graft may be low (< 1%) until 6 weeks after transplant if limiting doses of culture-initiating HSPCs were transplanted. Too low number of circulating human cells may decrease the reliability of further clonal analysis.

76 For each mouse, collect 100-200 µL of peripheral blood from tail vein sampling.

77 Homogenized well by repeatedly flicking the Eppendorf tubes.

78 Transfer 50 µL of blood in an individual new FACS tube for each mouse, add the Fc block anti-mouse (0.5 µL/sample) and anti-human (1 µL/sample), and incubate for 10 min at room temperature.

79 Add the dedicated anti-human antibodies for cell lineages characterization (see Reagent setup) and incubate for 15 min at 4°C.

80 Add 50 µL of FBS and vortex well before to perform automated red blood cell lysis with the TQ-Prep™ Workstation following manufacturer's procedure.

81 Wash the cells by filling the tube with DPBS + 2% FBS and pellet cells by using the centrifuge 5810R (1100 rpm, room temperature, 10 min). Discard the supernatant.

PAUSE STEP – *The stained sample may be kept at 4°C for up to 1-2 hours.*

82 Add 1 µL of 7-AAD to each sample for live/dead staining and briefly mix by vortexing.

83 Incubate for 10 min and then perform flow cytometry analysis with BD FACS Canto II.

? TROUBLESHOOTING

Sample preparation for BAR-Seq analysis on peripheral blood of transplanted NSG mice

TIMING – **2 hours / time point**

84 Add 200 µL of ACK lysing buffer to the remaining 100-150 µL of blood sample and thoroughly vortex.

85 Incubate for 5 min at room temperature.

86 Wash with 500 µL of DPBS + 2% FBS and centrifuge (1100 rpm, room temperature, 10 min). Aspirate and discard the supernatant.

PAUSE STEP – *The cell pellet may be frozen at -80°C for long-term storage.*

87 Extract gDNA using the QIAamp DNA Micro Kit, accordingly to manufacturer's instructions. Use a NanoDrop spectrophotometer to quantify DNA concentration.

PAUSE STEP – *gDNA samples are now ready for BAR-Seq clonal analysis following steps from 104 to the end. gDNA can be frozen at +4 or -20°C for short- or long-term storage, respectively.*

Sample preparation for cell sorting from hematopoietic organs of transplanted NSG mice

TIMING – **3 hours / 10 animals**

88 After 18-20 weeks after transplant, NSG mice may be euthanized accordingly to the approved institution protocol.

89 Collect cells from spleen by crushing and from posterior legs' bone marrow by flushing.

Spleen:

CRITICAL STEP – In order to minimize clamps formation, we strongly recommend working on ice and using cold reagents when processing the spleen.

90 Crush the spleen and filter cells using the 40 µm cell strainer with cold MACS Buffer.

91 Centrifuge the homogenate (1300 rpm, 4 °C, 10 min) and then discard the supernatant.

92 Add 1 mL of cold ACK lysing buffer to the cell pellet, thoroughly vortex and incubate for 5 min at room temperature.

93 Wash with cold MACS buffer up to filling the tube, centrifuge (1300 rpm, 4°C, 10 min) and then discard the supernatant.

94 Resuspend the cell pellet with 5 mL of cold MACS Buffer and then filter cell suspension with the 40 µM cell strainer. Centrifuge (1300 rpm, 4°C, 10 min) and then discard the supernatant. Proceed to step 97.

Bone marrow:

95 Flush bone marrow from posterior legs using cold MACS buffer and a 10 mL syringe with 1 mL needle.

96 Filter the cells using a 40 µM cell strainer, centrifuge (1300 rpm, 4°C, 10 min) and then discard the supernatant. Proceed to step 97.

Staining, cell sorting from hematopoietic organs of transplanted NSG mice and sample preparation for BAR-Seq clonal analysis TIMING – **5 hours / 10 animals**

97 Resuspend the cell pellet with 200 µL of cold MACS Buffer, add the Fc block anti-mouse (1 µL/sample) and anti-human (2 µL/sample) and incubate for 10 min at room temperature.

98 Add the dedicated anti-human antibodies for cell lineages characterization (see Reagent setup) and incubate for 15 min at 4°C.

99 Wash the cells by filling the tube with DPBS + 2% FBS and pellet cells by using the centrifuge 5810R (1100 rpm, room temperature, 10 min). Discard the supernatant.

PAUSE STEP – *The stained sample may be kept at 4°C for up to 1-2 hours.*

100 Resuspend the cells in the desired volume for cell sorting. Add 2 µL of 7-AAD to each sample for live/dead staining and briefly mix by vortexing.

CRITICAL STEP – *To facilitate high-speed sorting and to prevent clogging of the nozzle, filter the samples through 35 um filter immediately prior to sorting and dilute them such that at a flow rate of 2.0 an event rate of 5,000/10,000 events/second is not exceeded.*

101 Sort cell subpopulations of interest with BD FACSAria Fusion and collect samples in 1.5 mL Eppendorf tubes containing 500 µL of DPBS.

CRITICAL STEP – *We recommend using unstained and single-stained controls to set up compensation. Rainbow beads (SPHERO™ Rainbow Calibration Particles) should be included to standardize the experiments and have to be run before each acquisition.*

102 Centrifuge the sorted cells (2250 rpm, room temperature, 10 min). Aspirate and discard the supernatant.

CRITICAL STEP – *We recommend not to perform this step if the number of sorted cells is below 1,000 events and instead directly freeze the cells or proceed with gDNA extraction.*

PAUSE STEP – *The cell pellet can be frozen at -80°C for long-term storage.*

103 Extract gDNA using the QIAamp DNA Micro Kit, accordingly to manufacturer's instructions. Use a NanoDrop spectrophotometer to quantify DNA concentration.

PAUSE STEP – *gDNA samples are now ready for BAR-Seq clonal analysis following steps from 104 to the end. gDNA may be frozen at +4 or -20°C for short- or long-term storage, respectively.*

BAR-Seq library preparation TIMING – **4 hours / 49 samples**

CRITICAL STEP – *We advise performing steps 104-105 using LoBind disposable material and working in dedicated vertical laminal-flow hood, unless otherwise specified. We also recommend to carefully clean pipettes and the working area with DNA decontamination reagents before moving to the next steps.*

104 Set up the first PCR reaction for samples from step 74, 87 and/or 103 (up to a maximum of 49 samples + 1 "no template control") on ice in a LoBind 96-well plate.

Component	Amount (each reaction)
gDNA (from step 48, 74, 87 or 103)	100 ng (or up to 10-15 µL)
5X GoTaq® reaction buffer	5 µL
$MgCl_2$ (25 mM)	4 µL
dNTPs (10 mM)	1.2 µL
Primer PCR1_Fw (100 µM)	0.5 µL
Primer PCR1_Rv (100 µM)	0.5 µL
GoTaq G2 DNA Polymerase (5U/µl)	0.5 µL
Molecular-grade H_2O	Up to 50 µL

Seal the plate with a pierceable foil, briefly mix and spin. Load the PCR reactions on a thermocycler machine using the following PCR program.

Step	Number of cycles	Temperature	Duration
1	1	95°C	5 min
2	Up to 15-20	95°C	30 sec

		Tm – 5°C	30 sec
		72°C	depending on amplicon length
3	1	72°C	5 min
4	1	4°C	-

CRITICAL STEP – *Pfu DNA Polymerase (Promega) can be also used to further increase fidelity during amplification. In this case, the PCR program should be modified accordingly to manufacturer's instructions.*

CRITICAL STEP – *Plasmid DNA from step 47 can be used instead of gDNA to characterize library diversity.*

105Set up the second PCR reaction on ice in a new LoBind 96-well plate.

Component	Amount (each reaction)
5X GoTaq® reaction buffer	5 µL
$MgCl_2$ (25 mM)	4 µL
dNTPs (10 mM)	1.2 µL
GoTaq G2 DNA Polymerase	0.5 µL
Molecular-grade H_2O	Up to 44 µL

In a vertical laminal-flow hood, carefully add 0.5 µL of primer "PCR2_Fw#" and 0.5 µL of primer "PCR2_Rv#" for each sample. In a dedicated bench, carefully add 5 µL of PCR product from step 104 by piercing the foil with the tips loaded on a multichannel pipette. Seal the plate with a new pierceable foil, briefly mix and spin. Load the PCR reactions on a thermocycler machine using the following PCR program.

Step	Number of cycles	Temperature	Duration
1	1	95°C	5 min
2	Up to 15-20	95°C	30 sec
		Tm – 5°C	30 sec
		72°C	depending on amplicon length

3	1	72°C	5 min
4	1	4°C	-

CRITICAL STEP – *Pfu DNA Polymerase (Promega) can be also used to further increase fidelity during amplification. In this case, the PCR program should be modified accordingly to manufacturer's instructions.*

! CAUTION – *Addition of primers and PCR products should be performed very carefully to avoid cross-contaminations among samples.*

106 Equilibrate at room temperature the High Sensitivity D1000 Reagents and ScreenTape for 20 min.

107 Vortex High Sensitivity D1000 Reagents for 30 sec.

108 Separately prepare 1:10 dilutions of 1 µL PCR products from the second amplification round. Mix by briefly vortexing.

109 To verify for successful amplification, load on Agilent 4200 TapeStation the diluted PCR products following manufacturer's instructions.

PAUSE STEP – *PCR products can be frozen at -20°C before purification.*

? TROUBLESHOOTING

110 Purify the PCR product using AMPure XP beads following manufacturer's instructions (final volume = 17.5 µL).

111 Quantify amplicon concentration for each sample by Agilent 4200 TapeStation following steps 106-109 and manufacturer's instructions.

PAUSE STEP – *Purified amplicons can be frozen at -20°C for long-term storage.*

Next-generation sequencing of the BAR-Seq amplicons TIMING – **1 day / sequencing run**

112 Submit samples to your sequencing facility. We tested the protocol on both Illumina NextSeq500 and MiSeq platforms, following the manufacturer's instructions. Single-end sequencing is sufficient.

CRITICAL STEP – *We recommend adding PhiX at 30-40% to ensure proper reading of amplicons.*

BAR-Seq bioinformatic analyses TIMING – **8 hours / 49 samples**

Local execution

113 (Optional) Pre-process the input reads (usually in FASTQ format) to check their quality and perform a filter/trim to remove low quality portions or sequencing adapters (i.e. using FastQC and Trimmonatic), if required.

114 Extract the barcode sequences from the input reads, based on the amplicon structure, using TagDust2 (**Figure 3**).

CRITICAL STEP – *The percentage of reads from which a valid BAR is extracted should be above 80%.*

? TROUBLESHOOTING

115 Perform the filtering of extracted reads based on their length and structure, and then correct the erroneous ones by employing the graph-based procedure which identifies and merges the ego-(sub)networks used to represent the similarity among barcode sequences.

116 Select the threshold for the minimum BAR abundance (e.g. 3) and discard those having lower counts. The remaining sequences are the set of valid BARs.

? TROUBLESHOOTING

117 Rank the BARs from the most to the least abundant and identify the dominant and rare (sub)populations by applying a saturation-based approach based on the selected threshold (e.g. 90%).

118 (Optional) If more samples are analyzed, compute the BAR sharing among them, by considering the overall set of valid sequences and keeping track of their abundance in each sample. The same can be done by considering only the dominant BARs.

119 (Optional) Compute the sample diversity and the level of richness for each sample to check if the set of BARs is informative enough. Optimal experiments should achieve high richness values (R>0.7), while low richness levels (R<0.3) could indicate poor sequencing depth. In this case, amplicon preparation and/or sequencing must be re-performed by following the suggestions reported in the **Troubleshooting** guide.

Online execution (alternative to steps 113 – 119)

120 Run the BAR-Seq pipeline online by visiting the web-application page http://www.bioinfotiget.it/barseq (**Figure 4**).

121 Upload your set of pre-processed FASTQ files (one for each sample) containing the input sequences.

CRITICAL STEP – *We set a limit of 500MB for the total size of the uploaded samples to avoid overloading the server. Anyway, upon specific request, this size limit can be increased.*

122 Specify the known sequences flanking the BAR within the amplicon (upstream and downstream).

123 Select filter thresholds for "Min. Barcode Count" and "Saturation Threshold (%)".

124 (Optional) In the "Advanced Options" select the "Structural Filter" by choosing among "No Filter", "Filter Nucleotide with Freq. < 1%", and "Fixed Structure" which requires an IUPAC sequence with the structure. The "Graph Edit Distance" can also be customized.

125 Press the "Run" button to perform the computation remotely. BAR-Seq results are shown and can be downloaded.

CRITICAL STEP – *We advise to pay attention when specifying the two sequences flanking the BAR for its extraction. The percentage of reads from which a valid BAR is extracted should be above 80%.*

? TROUBLESHOOTING

126 (Optional) To check if the set of BARs is informative, compute the sample diversity and the level of richness for each sample by clicking on the "Check Diversity/Richness" button in the "Results" page. Select the desired confidence level and press "Run". Optimal experiments should achieve high richness values (R>0.7), while low richness levels (R<0.3) could indicate poor sequencing depth. If the set of BARs is not informative enough, amplicon preparation or sequencing must be re-performed by following the suggestions reported in the **Troubleshooting** guide.

CRITICAL STEP – *Alternatively, the downloaded text file containing the set of valid BARs (from step 125) can be manually uploaded by clicking on the "Check Diversity/Richness" tab in the homepage.*

CRITICAL STEP – *When analyzing the original plasmid/viral library, the table in the "Check Diversity/Richness" result page provides an estimation of the number of uniquely labeled cells that can be tracked with a certain confidence level by using the given library.*

? TROUBLESHOOTING

TIMING

Step 1-48, preparation of the barcoded HDR template: 4-5 days.

Step 49-55, thawing of CB CD34$^+$ HSPCs: 30 min hands-on, 3 days of culture.

Step 56-63, gene editing procedure of cultured HSPCs: 1-2 hours hands-on, 1 day of culture.

Step 64-71, transplantation of edited HSPCs in immunodeficient NSG mice: 2 hours hands-on.

Step 72-74, sample preparation for BAR-Seq analysis on cultured HSPCs: 2 hours hands on.

Step 75-83, phenotypic analysis of peripheral blood samples from transplanted NSG mice: 3 hours hands-on / time point.

Step 84-87, sample preparation for BAR-Seq analysis on peripheral blood of transplanted NSG mice: 2 hours hands-on / time point.

Step 88-96, sample preparation for cell sorting from hematopoietic organs of transplanted NSG mice: 3 hours hands-on / 10 animals.

Step 97-103, staining, cell sorting from hematopoietic organs of transplanted NSG mice and sample preparation for BAR-Seq clonal analysis: 5 hours hands-on / 10 animals.

Step 104-111, BAR-Seq library preparation: 4 hours hands-on / 49 samples.

Step 112, Next-generation sequencing of the BAR-Seq amplicons: 1 day / sequencing run

Step 113-119 or 120-126, BAR-Seq bioinformatics analysis: 8 hours / 49 samples.

TROUBLESHOOTING

Step	Problem	Possible reasons	Possible solutions
10	Absence of the peaks corresponding to successful digestion	Inefficient ssODN amplification	Decrease the amount of ssODN/reaction
			Repeat amplification using different polymerases (e.g. Pfu)
		Inefficient restriction enzyme activity (<40% digested product)	Increase the restriction enzyme units and/or the incubation time
			Use new stocks of restriction enzymes
17	Absence of the expected digestion pattern	Inefficient restriction enzyme activity	Increase the restriction enzyme units
			Use new stocks of restriction enzymes
		Ethanol traces in the plasmid prep	Precipitate the plasmid prep, wash, resuspend the pellet and perform again digestion
31	Too low number of colonies/plate	Poor transformation efficiency	Improve transformation efficiency by testing other ligation/transformation conditions
			Performing again backbone/insert digestions
	Too high number of colonies/plate	Highly efficient transformation	Use >100 plates/vial of ultracompetent cells in step 42

	High number of colonies in the "backbone only" control plates	Inefficient/incomplete digestion of the plasmid backbone	Repeat plasmid digestion increasing the units of restriction enzymes or decreasing the starting amount of plasmid
47/48 (related to Box1 and Box2)	Low complexity of the barcoded library	Inefficient ligation/transformation	Increase the number of ultracompetent cells vials used for transformation and/or the number of ligation reactions
			See troubleshooting for step 31
71	Low HDR editing efficiency	Inefficient gRNA cleavage activity	Evaluate cutting efficiency by T7 endonuclease assay, increase the RNP dose or change the Cas9:gRNA ratio
		Low permissiveness of the locus to HDR	Include Ad5-E4orf6/7 (+/- GSE56) mRNA during electroporation (note: GSE56 enhances clonality of edited HSPCs)
83	Low human cell engraftment	High toxicity of the editing procedure	Optimize editing conditions by decreasing AAV or RNP doses
			Include GSE56 mRNA during electroporation (note: GSE56 enhances clonality of edited HSPCs)
		HSPC donor variability	Pool HSPCs from different donors
109	Absent amplicon peak or low amplicon concentration	Low yield of gDNA (<1 ng/uL)	Start with >1,000 cells
		Low HDR editing efficiency (<1%) and or low human cell engraftment (<1%)	See troubleshooting for step 71/83
114 /116 /125	Low percentage of barcode retrieved	Possible errors in defining the read structure	Check barcode flanking sequences provided to TagDust in the BAR-Seq software utility
119 /126		Low number of reads to capture the barcode	Perform again amplicon preparation increasing the amount of DNA

	Poor richness for some samples	complexity of the sample	Re-sequence samples with poor richness increasing depth (i.e. decreasing the number of samples multiplexed in the run

ANTICIPATED RESULTS

BAR-Seq pipeline is designed to perform clonal tracking analyses of edited cells both *in vitro* and *in vivo*, without any limitation due to the locus, donor template or nuclease platform. We easily obtain BAR-Seq plasmid libraries with 3-8x10^5 unique BARs and we observe only minimal skewing in BAR abundance from plasmid to viral library[9]. As discussed above, the required library complexity and diversity strictly depend on the expected complexity of the cell population of interest and the desired number of traceable cells. The BAR-Seq web application provides a function which estimates whether the sequenced plasmid/viral library is suitable for clonal tracking of a desired number of cells. Software outputs for good (highly diverse) and bad (poorly diverse) libraries are shown in **Figure 5a**.

The experimental procedure described in this Protocol allows efficient BAR extraction from gDNA of edited cells. An example of the amplicon library profile is shown in **Figure 5b**.

When applied to clonal tracking of edited HSPCs, BAR-Seq uncovered the multilineage and self-renewing capacity of engrafting HDR-edited HSPCs in human hematochimeric mice with the majority of clones shared among cell lineages long-term after transplant. Moreover, the human HDR-edited cell graft was composed by few highly abundant clones. Higher complexity of the clonal repertoire can be achieved by dampening the editing-induced p53-dependent DNA-damage response and enhancing editing efficiency[9].

DATA AVAILABILITY

The BAR-Seq software with some example datasets are provided as a zip file in the **Supplementary Materials**. Moreover, the datasets available in Gene Expression Omnibus with the access code GSE144340 have been analyzed using the BAR-Seq pipeline as described in[9], including those presented as examples in **Figure 4**.

CODE AVAILABILITY

The scripts for BAR-Seq analysis are freely available at https://bitbucket.org/bereste/bar-seq. The BAR-Seq webtool is freely available at http://www.bioinfotiget.it/barseq.

ETHICAL STATEMENT

All experiments and procedures involving animals were performed with the approval of the Animal Care and Use Committee of the San Raffaele Hospital (IACUC no. 749) and authorized by the Italian Ministry of Health and local authorities accordingly to Italian law.

The use of human HSPCs was approved by the Ospedale San Raffaele Scientific Institute Bioethical Committee (TIGET-HPCT protocol).

REFERENCES

1. Naldini, L. Gene therapy returns to centre stage. *Nature* (2015). doi:10.1038/nature15818
2. Bramlett, C. *et al.* Clonal tracking using embedded viral barcoding and high-throughput sequencing. *Nat. Protoc.* (2020). doi:10.1038/s41596-019-0290-z
3. Cavazzana-Calvo, M. *et al.* Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science (80-.).* **288**, 669–672 (2000).
4. Biffi, A. *et al.* Lentiviral hematopoietic stem cell gene therapy benefits metachromatic leukodystrophy. *Science (80-.).* **341**, (2013).
5. Aiuti, A. *et al.* Lentiviral hematopoietic stem cell gene therapy in patients with wiskott-aldrich syndrome. *Science (80-.).* **341**, (2013).
6. Scala, S. *et al.* Dynamics of genetically engineered hematopoietic stem and progenitor cells after autologous transplantation in humans. *Nat. Med.* (2018). doi:10.1038/s41591-018-0195-3
7. Wu, C. *et al.* Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. *Cell Stem Cell* (2014). doi:10.1016/j.stem.2014.01.020
8. Doudna, J. A. The promise and challenge of therapeutic genome editing. *Nature* **578**, 229–236 (2020).
9. Ferrari, S. *et al.* Efficient gene editing of human long-term hematopoietic stem cells validated by clonal tracking. *Nat. Biotechnol.* (2020). doi:10.1038/s41587-020-0551-y
10. Kebschull, J. M. & Zador, A. M. Cellular barcoding: lineage tracing, screening and beyond. *Nature Methods* (2018). doi:10.1038/s41592-018-0185-x
11. Lu, R., Neff, N. F., Quake, S. R. & Weissman, I. L. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat. Biotechnol.* (2011). doi:10.1038/nbt.1977
12. San Filippo, J., Sung, P. & Klein, H. Mechanism of Eukaryotic Homologous Recombination. *Annu. Rev. Biochem.* **77**, 229–257 (2008).
13. Yeh, C. D., Richardson, C. D. & Corn, J. E. Advances in genome editing through control of

DNA repair pathways. *Nat. Cell Biol.* **21**, 1468–1478 (2019).

14. Wang, J. *et al.* Homology-driven genome editing in hematopoietic stem and progenitor cells using ZFN mRNA and AAV6 donors. *Nat. Biotechnol.* **33**, 1256–1263 (2015).

15. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science (80-.).* (2020). doi:10.1126/science.aaw3381

16. Wagenblast, E. *et al.* Functional profiling of single CRISPR/Cas9-edited human long-term hematopoietic stem cells. *Nat. Commun.* **10**, (2019).

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

18. Genovese, P. *et al.* Targeted genome editing in human repopulating haematopoietic stem cells. *Nature* **510**, 235–240 (2014).

19. Dever, D. P. *et al.* CRISPR/Cas9 β-globin gene targeting in human haematopoietic stem cells. *Nature* **539**, 384–389 (2016).

20. McKenna, A. *et al.* Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science (80-.).* (2016). doi:10.1126/science.aaf7907

21. Kalhor, R. *et al.* Developmental barcoding of whole mouse via homing CRISPR. *Science (80-.).* (2018). doi:10.1126/science.aat9804

22. van Overbeek, M. *et al.* DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Mol. Cell* (2016). doi:10.1016/j.molcel.2016.06.037

23. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4192

24. Rago, C., Vogelstein, B. & Bunz, F. Genetic knockouts and knockins in human somatic cells. *Nat. Protoc.* (2007). doi:10.1038/nprot.2007.408

25. Khan, I. F., Hirata, R. K. & Russell, D. W. AAV-mediated gene targeting methods for human cells. *Nat. Protoc.* (2011). doi:10.1038/nprot.2011.301

26. Bak, R. O., Dever, D. P. & Porteus, M. H. CRISPR/Cas9 genome editing in human hematopoietic stem cells. *Nat. Protoc.* (2018). doi:10.1038/nprot.2017.143

27. Naik, S. H., Schumacher, T. N. & Perié, L. Cellular barcoding: A technical appraisal. *Experimental Hematology* **42**, 598–608 (2014).

28. Bystrykh, L. V. & Belderbos, M. E. Clonal analysis of cells with cellular barcoding: When numbers and sizes matter. in *Methods in Molecular Biology* (2016). doi:10.1007/7651_2016_343

29. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* (2009). doi:10.1038/nmeth.1318
30. Roth, T. L. *et al.* Reprogramming human T cell function and specificity with non-viral genome targeting. *Nature* (2018). doi:10.1038/s41586-018-0326-5
31. Gurumurthy, C. B. *et al.* Creation of CRISPR-based germline-genome-engineered mice without ex vivo handling of zygotes by i-GONAD. *Nat. Protoc.* (2019). doi:10.1038/s41596-019-0187-x
32. Tiscornia, G., Singer, O. & Verma, I. M. Production and purification of lentiviral vectors. *Nat. Protoc.* (2006). doi:10.1038/nprot.2006.37
33. Galibert, L. & Merten, O.-W. Latest developments in the large-scale production of adeno-associated virus vectors in insect cells toward the treatment of neuromuscular diseases. *J. Invertebr. Pathol.* **107 Suppl**, S80–S93 (2011).
34. Grieger, J. C., Choi, V. W. & Samulski, R. J. Production and characterization of adeno-associated viral vectors. *Nature Protocols* (2006). doi:10.1038/nprot.2006.207
35. Doria, M., Ferrara, A. & Auricchio, A. AAV2/8 vectors purified from culture medium with a simple and rapid protocol transduce murine liver, muscle, and retina efficiently. *Hum. Gene Ther. Methods* (2013). doi:10.1089/hgtb.2013.155
36. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* (2013). doi:10.1038/nprot.2013.143
37. Boitano, A. E. *et al.* Aryl hydrocarbon receptor antagonists promote the expansion of human hematopoietic stem cells. *Science (80-.).* **329**, 1345–1348 (2010).
38. Fares, I. *et al.* Pyrimidoindole derivatives are agonists of human hematopoietic stem cell self-renewal. *Science (80-.).* **345**, 1509–1512 (2014).
39. Schiroli, G. *et al.* Preclinical modeling highlights the therapeutic potential of hematopoietic stem cell gene editing for correction of SCID-X1. *Sci. Transl. Med.* **9**, (2017).
40. Lux, C. T. *et al.* TALEN-Mediated Gene Editing of HBG in Human Hematopoietic Stem Cells Leads to Therapeutic Fetal Hemoglobin Induction. *Mol. Ther. - Methods Clin. Dev.* (2019). doi:10.1016/j.omtm.2018.12.008
41. Cromer, M. K. *et al.* Global Transcriptional Response to CRISPR/Cas9-AAV6-Based Genome Editing in CD34+ Hematopoietic Stem and Progenitor Cells. *Mol. Ther.* (2018). doi:10.1016/j.ymthe.2018.06.002
42. Schiroli, G. *et al.* Precise Gene Editing Preserves Hematopoietic Stem Cell Function following Transient p53-Mediated DNA Damage Response. *Cell Stem Cell* **24**, 551-565.e8 (2019).

43. De Ravin, S. S. *et al.* CRISPR-Cas9 gene repair of hematopoietic stem cells from patients with X-linked chronic granulomatous disease. *Sci. Transl. Med.* **9**, (2017).
44. Petrillo, C. *et al.* Cyclosporine H Overcomes Innate Immune Restrictions to Improve Lentiviral Transduction and Gene Editing In Human Hematopoietic Stem Cells. *Cell Stem Cell* **23**, 820-832.e9 (2018).
45. Clement, K. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nature Biotechnology* **37**, 224–226 (2019).
46. Brinkman, E. K., Chen, T., Amendola, M. & Van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* (2014). doi:10.1093/nar/gku936
47. Magoč, T. & Salzberg, S. L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btr507
48. Lassmann, T. TagDust2: A generic method to extract reads from sequencing data. *BMC Bioinformatics* **16**, 24 (2015).
49. Thielecke, L., Cornils, K. & Glauche, I. GenBaRcode: A comprehensive R-package for genetic barcode analysis. *Bioinformatics* (2020). doi:10.1093/bioinformatics/btz872
50. Del Core Luca, Montini Eugenio, Di Serio Clelia, C. A. Dealing with Data Evolution and Data Integration: An approach using Rarefaction. in *49th Scientific meeting of the Italian Statistical Society* (2018).

ACKNOWLEDGEMENTS

We thank all members of LN's laboratory for discussion, the IRCCS San Raffaele Hospital Flow Cytometry facility (FRACTAL), the IRCCS San Raffaele Center for Omics Sciences (COSR), A. Auricchio and M. Doria (Telethon Institute of GEnetics and Medicine, the TIGEM Vector Core, Pozzuoli, Italy) for providing AAV6 vectors, L. Periè (Institute Curie, Paris, France) and J. Urbanus (the Netherlands Cancer Institute, Amsterdam, the Netherlands) for advice on the BAR cloning strategy, G. Schiroli (SR-Tiget) for initial help with the design of the BAR-Seq strategy, D. Lazarevic (COSR) for the help with BAR-Seq amplicon sequencing and A. Calabria (SR-Tiget) for the help with richness estimation. This work was supported by grants to: LN from Telethon (TIGET grant E4), the Italian Ministry of Health (PE-2016-02363691; E-Rare-3 JTC 2017), the Italian Ministry of University and Research (PRIN 2017 Prot. 20175XHBPN), the EU Horizon 2020 Program (UPGRADE), and from the Louis-Jeantet Foundation through the 2019 Jeantet-Collen Prize for Translational Medi cine; PG from Telethon (TIGET grant E3) and the Italian Ministry of Health (GR-2013-02358956). SF conducted this study as partial fulfillment of their Ph.D. in Molecular Medicine,

International Ph.D. School, Vita-Salute San Raffaele University (Milan, Italy). AJ conducted this study as partial fulfillment of his Ph.D. in Translational and Molecular Medicine - DIMET, Milano-Bicocca University (Monza, Italy).

AUTHOR CONTRIBUTIONS

S.F., P.G., and L.N conceived and developed the protocol. S.B., D.C., I.M. developed the BAR-Seq bioinformatic pipeline. S.B. and I.M. developed the online BAR-Seq tool. S.F., A.J., A.L. and P.G. developed and optimized the gene editing protocol. S.F., S.B., A.J., P.G., I. M., L.N. wrote the manuscript.

COMPETING INTEREST

L.N. and P.G. are inventors of patents on applications of gene editing in HSPCs owned and managed by the San Raffaele Scientific Institute and the Telethon Foundation, including a patent application on improved gene editing filed by S.F., A.J., P.G. and L.N. L.N. is founder and quota holder and P.G. is quota holder of GeneSpire, a startup company aiming to develop *ex vivo* gene editing in genetic diseases. All other authors declare no conflict of interest.

FIGURE LEGENDS

Figure 1. BAR-Seq pipeline for clonal tracking of edited cells. a, Schematics of the gene editing outcomes upon nuclease-induced DNA double strand break. BAR-Seq allows clonal tracking of cells edited by HDR. **b,** Experimental workflow of the BAR-Seq pipeline.

Figure 2. Alternative configurations and generation of the BAR-Seq HDR template library. a, Schematics of the BAR-Seq HDR template showing the possible configurations for BAR cloning, either outside (top) or inside (bottom) of an expression cassette. In the latter configuration, BAR will be transcribed with the gene of interest. Red arrows indicate primer binding sites for the assessment of BAR-Seq library diversity. **b,** Alternative PCR configurations for BAR amplification after integration. The “In-Out” PCR approach exclusively amplifies on-target HDR-integrated BARs. Red arrows indicate primer binding sites for the generation of BAR-Seq amplicon from edited cells. **c,** Schematic representation of the experimental procedure and expected results for the BAR-Seq library generation (steps 1-46).

Figure 3. BAR-Seq bioinformatic pipeline. Workflow of the BAR-Seq pipeline with representative results.

Figure 4. Web application for BAR-Seq bioinformatic analyses. Graphic interface of the homepage with demo samples uploaded (top left) and of the result page (right) reporting also sharing as heatmap (bottom left).

Figure 5. Anticipated results of the BAR-Seq pipeline. a, Expected outputs of the BAR-Seq web application upon submission of sequencing data of highly diverse (top) or poorly diverse (bottom) plasmid/viral libraries. **b,** Agilent 4200 traces of two amplicon library preparations.

BOX 1: Diversity assessment of the BAR-Seq plasmid library

EXTRA-REAGENTS

- Primers (Sigma-Aldrich, Metabion or another vendor; see **Supplementary Table 3** for ordering details)

PROCEDURE

TIMING – **3 days**

1 Amplify by PCR the BAR region from the plasmid library, purify the amplicon product and deep sequence by following steps from 104 to 112. For the first PCR step, primers ("Vector_PCR1"; see **Supplementary Table 3**) must be both designed to bind the viral library sequence and to flank the BAR. For the second PCR step, primers ("Vector_PCR2") are listed in **Supplementary Table 3**.

2 Process sequencing data by following steps from 120 to 126 (alternatively from 113 to 119).

BOX 2: Diversity assessment of the BAR-Seq viral library

EXTRA-REAGENTS

- DpnI (New England Biolabs, cat. no. R0176L)
- DNeasy Blood and Tissue Kit (QIAGEN, cat no. 69504)
- Primers (Sigma-Aldrich, Metabion or another vendor; see **Supplementary Table 3** for ordering details)

EXTRA-BIOLOGICAL MATERIAL

- K-562 cell line (ATCC, cat. no. CCL-243, RRID:CVCL_0004) https://scicrunch.org/resolver/RRID:CVCL_0004

  ! CAUTION – *The cell lines used in your research should be regularly checked to ensure they are authentic and are not infected with mycoplasma.*

PROCEDURE

TIMING – **4 days**

! CAUTION – *This procedure must be performed in sterile hood.*

CRITICAL STEP – *Cell lines highly permissive to transduction with the viral vector of interest (e.g K-562) should be used to assess the diversity of the BAR-Seq library.*

1 Transduce 1-10x10^6 K-562 cells with an appropriate volume of the BAR-Seq viral library from step 48. We advise to transduce cells at high multiplicity of infection (MOI = 100 transducing units/mL for IDLV; MOI = 10^4 for AAV). Incubate transduced cells for 24 h at 37°C in a 5% CO_2 and 20% O_2 humidified atmosphere.

CRITICAL STEP – *The number of transduced cells and the MOI are critical parameters in order to exhaustively sequence the BAR-Seq viral library and may vary accordingly to the expected library complexity and diversity. To be conservative, the values given above are suitable for libraries with expected complexity <10^6.*

2 Collect transduced cells in a suitable sterile Falcon tube, add 10 volumes of DPBS and pellet them using the centrifuge 5810R (1100 rpm, room temperature, 10 min).

PAUSE STEP – *The cell pellet may be frozen at -80°C for long-term storage.*

3 Extract the DNA using the DNeasy Blood and Tissue Kit, accordingly to manufacturer's instructions. Use a NanoDrop spectrophotometer to quantify DNA concentration.

4 (Optional) Set up digestion reactions of the DNA from the previous step with the methylation-sensitive restriction enzyme DpnI to cleave and drop out residual plasmid contaminants.

Component	Amount
DNA	5 µL
10X CutSmart buffer	5 µL
DpnI	2 µL
Molecular-grade H_2O	38 µL

PAUSE STEP – *DNA can be frozen at -20°C for storage.*

5 Amplify by PCR the BAR region, purify the amplicon product and deep sequence by following steps from 104 to 112. For the first PCR step, primers ("Vector_PCR1"; see **Supplementary Table 3**) must be both designed to bind the viral library sequence and to flank the BAR. For the second PCR step, primers ("Vector_PCR2") are listed in **Supplementary Table 3**.

6 Process sequencing data by following steps from 120 to 126 (alternatively from 113 to 119).

BOX 3: Clonogenic *in vitro* assay of edited HSPCs:

EXTRA-REAGENTS

- MethoCult H4434 Classic (StemCell Technologies, cat. no. 04434)

PROCEDURE

TIMING – day '+4' of the editing protocol – **1 hour**

! CAUTION – *This procedure must be performed in sterile hood.*

1 Count the number of viable HSPCs.

2 For each experimental condition, prepare 6 mL of methylcellulose-based medium supplemented with 100 IU/mL penicillin, 100 µg/mL streptomycin, 2% (vol/vol) glutamine in a 15 mL Falcon tube.

3 For each experimental condition, collect 2,400 cells and add them to the supplemented medium from step 2. Thoroughly vortex and incubate 10 min at room temperature.

4 For each experimental condition, seed three wells (triplicate) of P6-well plate with 1.5 mL of the medium from step 3 (~600 cells/well).

CRITICAL STEP – *Reverse pipetting technique is recommended due to viscosity of methylcellulose media and to avoid bubbles formation. We also advise to pipette 1.5 mL by pipetting two times 750 µL using the L-1000XLS+ pipette.*

5 Incubate for **14 days** at 37°C in a 5% CO_2 and 20% O_2 humidified atmosphere.

6 Count the number of colonies in each well.

BOX 4: Phenotypic characterization of edited HSPCs *in vitro*:

EXTRA-REAGENTS

- Anti-human CD34-PE-Cy7, Clone 8G12 (BD Biosciences, cat. no. 348811, RRID: AB_2868855) https://scicrunch.org/resolver/RRID:AB_2868855
- Anti-human CD133/2-PE, Clone REA816 (Miltenyi Biotec, cat. no. 130-112-157, RRID: AB_2654900) https://scicrunch.org/resolver/RRID:AB_2654900

- Anti-human CD90-APC, Clone 5E10 (BD Biosciences, cat. no. 559869, RRID: AB_398677) https://scicrunch.org/resolver/RRID:AB_398677
- 7-AAD Viability Staining Solution (BioLegend, cat. no. 420404)

PROCEDURE

TIMING – day '+7' of the editing protocol – **2 hours**

1 Collect >50,000 cultured HSPCs in an individual new FACS tube for each experimental condition.

2 Add the dedicated anti-human antibodies for cell phenotypic characterization and incubate for 15 min at 4°C.

Target	Fluorochrome	Working Dilution	Amount x 100µl (each reaction)
CD34$^+$	PE-Cy7	(1:40)	2.5 µL
CD133$^+$	PE	(1:40)	2.5 µL
CD90$^+$	APC	(1:65)	1.5 µL

3 Wash the cells by filling the tube with DPBS + 2% FBS and pellet them using the centrifuge 5810R (1100 rpm, room temperature, 10 min).

PAUSE STEP – *The stained sample may be kept at 4°C for up to 1-2 hours.*

4 Add 1 µL of 7-AAD to each sample for live/dead staining and briefly mix by vortexing.

5 Incubate for 10 min and then perform flowcytometry analysis with BD FACS Canto II.

BOX 5: Quantification of HDR editing efficiency in *in vitro* edited HSPCs:

EXTRA-REAGENTS

- TTC5 HEX (20X) PrimePCR ddPCR Copy Number Assay: TTCA, Human (Biorad, cat. no. dHSACP2506733)
- ddPCR Supermix for Probes (no dUTP) (Biorad, cat. no. 1863024)
- DG8 Cartridges for QX200/QX100 Droplet Generator (Biorad, cat. no. 1864008)
- Droplet generation oil for probes (Biorad, cat. no. 1863005)
- Eppendorf twin.tec PCR plate, 96 well, semi-skirted (Eppendorf, cat. no. 951020303)
- Custom primers and FAM probe specific for the HDR-edited locus (Sigma-Aldrich, Metabion or another vendor; primers and probes used in[9] for the *AAVS1* locus are listed below)

AAVS1 3' integration donor-genome junction	FW	GATTGGGAAGACAATAGCAG
	RV	TCTTGGGAAGTGTAAGGAAG
	Probe (FAM)	CCAGATAAGGAATCTGCCTA

EXTRA-EQUIPMENT

- BioRad Automated Droplet Generator (Bio-Rad, cat no. 1864101)
- BioRad PX1 PCR Plate Sealer (Bio-Rad, cat. no. 1814000)
- BioRad QX200 Droplet Reader (Bio-Rad, cat no. 1864003)
- QuantaSoft™ Analysis Pro software (Bio-Rad)

PROCEDURE

TIMING – day '+7' of the editing protocol – **6 hours**

1 Collect up to 50,000 cultured HSPCs for each experimental condition, add 10 volumes of DPBS and pellet them using the centrifuge 5430 (2250 rpm, room temperature, 10 min).

2 Aspirate and discard the supernatant.

PAUSE STEP – *The cell pellet can be frozen at -80°C for long-term storage.*

3 Extract gDNA using the QIAamp DNA Micro Kit, accordingly to manufacturer's instructions. Use a NanoDrop spectrophotometer to quantify DNA concentration.

PAUSE STEP – *gDNA can be frozen at +4 or -20°C for short- or long-term storage, respectively.*

4 Prepare PCR mix and add gDNA material in a semi-skirted PCR microplate.

Component	Amount (each reaction)
gDNA	5 ng (or up 5 µL)
ddPCR Supermix for Probes (no dUTP)	11 µL
ddPCR custom primers/probe FAM assay (20X)	1.1 µL
Human TTC5 HEX assay (20X)	1.1 µL
Molecular-grade H_2O	Up to 22 µL

! CAUTION – *This procedure must be performed in DNA/RNA-free hood.*

5 Proceed to droplet formation according to the manufacturer's instructions (Bio-Rad).

6 Seal the plate with a pierceable foil and load the PCR reactions on a thermocycler machine using the following PCR program.

Step	Number of cycles	Temperature	Duration
1	1	95°C	10 min
2	40	94°C	30 sec
		Tm – 55°C	1 min
		72°C	depending on amplicon length
3	1	98°C	10 min
4	1	4°C	-

7 Proceed to plate reading and analyze results according to the manufacturer's instructions (Bio-Rad).

CRITICAL STEP – *The percentage of alleles carrying HDR-mediated integration can be calculated as follow: no. of targeted locus$^+$ droplets / no. of TTC5$^+$ droplets x 100. If targeting a sexual chromosome in male cells, the result of the formula must be multiplied by 2 in order to fit with the TTC5 reference gene used for normalization, which is located in an autosomal chromosome. In this case, the percentage of alleles carrying HDR-mediated integration corresponds to the percentage of HDR-edited cells.*

BOX 6: Quantification of the NHEJ editing efficiency in *in vitro* edited HSPCs:

EXTRA-REAGENTS

- T7 Endonuclease I (New England Biolabs, cat. no. M0302L)
- Custom primers specific for the locus of interest (Sigma-Aldrich, Metabion or another vendor; primers used in[9] for the *AAVS1* locus are listed below)

NHEJ *AAVS1*	FW	GCCCTGGCCATTGTCACTTT
	RV	GGACTAGAAAGGTGAAGAGCC

PROCEDURE

TIMING – day '+7' of the editing protocol – **5 hours**

CRITICAL STEP – *Primers must be specific for the locus of interest and amplify a sequence between 500-1000 bp length.*

1 Set up a PCR mix as follow:

CRITICAL STEP – *gDNA material extracted for quantification of HDR efficiency can be used for this analysis.*

Component	Amount (each reaction)
gDNA (from Box 5)	50-100 ng
5X GoTaq® reaction buffer	5 µL
$MgCl_2$ (25 mM)	4 µL
dNTPs (10 mM)	1.2 µL
Primer Fw (100 µM)	0.5 µL
Primer Rv (100 µM)	0.5 µL
GoTaq G2 DNA Polymerase (5U/µl)	0.5 µL
Molecular-grade H_2O	Up to 50 µL

2 Seal the plate with a new pierceable foil, briefly mix and spin. Load the PCR reactions on a thermocycler machine using the following PCR program:

Step	Number of cycles	Temperature	Duration
1	1	95°C	10 min
2	40	95°C	45 sec
		Tm – 5°C	30 sec
		72°C	30 sec
3	1	72°C	10 min
4	1	4°C	-

3 Immediately run an annealing program as follow. The annealing step is crucial to generate sequence mismatches.

Step	Number of cycles	Temperature	Duration
1	1	95°C (-2°C/s)	10 min

2	1	85°C (-0.1°C/s)	1 min
3	1	75°C (-0.1°C/s)	1 min
4	1	65°C (-0.1°C/s)	1 min
5	1	55°C (-0.1°C/s)	1 min
6	1	45°C (-0.1°C/s)	1 min
7	1	35°C (-0.1°C/s)	1 min
8	1	25°C (-0.1°C/s)	1 min
9	1	4°C	-

4 Set up an enzymatic digestion of the PCR products with the T7 endonuclease as follow:

Component	Amount
PCR product	20 µL
10X NEB2 buffer	2,5 µL
T7	0,5 µL
Molecular-grade H_2O	2 µL

5 Incubate at 37°C for maximum 60 min. Do not exceed with the digestion time to avoid unspecific cleavage.

PAUSE STEP – *PCR products can be frozen at -20°C before purification.*

6 Equilibrate at room temperature the High Sensitivity D1000 Reagents and ScreenTape for 20 min.

7 Vortex High Sensitivity D1000 Reagents for 30 sec.

8 To verify for nuclease activity, load on Agilent 4200 TapeStation the digested PCR products following manufacturer's instructions.