Scalable tagging of endogenous genes by homology-independent intron targeting

Yevgeniy V Serebrenik\textsuperscript{1,2,*}, Stephanie E Sansbury\textsuperscript{1,2,*}, Saranya Santhosh Kumar\textsuperscript{1,2}, Jorge Henao-Mejia\textsuperscript{3,4,5} and Ophir Shalem\textsuperscript{1,2,†}

\textsuperscript{1}Center for Cellular and Molecular Therapeutics, Children’s Hospital of Philadelphia, Philadelphia, PA 19104, USA
\textsuperscript{2}Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
\textsuperscript{3}Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
\textsuperscript{4}Institute for Immunology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
\textsuperscript{5}Division of Protective Immunity, Department of Pathology and Laboratory Medicine, Children’s Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA 19104

* Equal contribution
† Correspondence: shalemo@upenn.edu

Abstract
Genome editing tools have simplified the generation of knock-in gene fusions, yet the requirement for gene-specific homology directed repair (HDR) templates still hinders the scalability of most approaches. Here, we combine intron-based protein trapping with homology independent repair-based editing and demonstrate precise and efficient gene tagging that can be easily scaled due to use of a generic donor. As editing is done in introns, this approach tolerates mutations in the unedited allele, disruptive indels, and allows for flexible donor and sgRNA design.

Main
Fusing endogenous proteins with fluorescence or epitope tags is a widely used and essential approach for studying proteins within their natural regulatory context. The advent of CRISPR tools for modifying the genome\textsuperscript{1-3} has made this easier and even more accessible, yet scalability is still very limited. The need for a gene-specific homology directed repair (HDR) template requires costly synthesis or labor-intensive molecular cloning, and since precise targeting must be achieved in frame with the coding sequence, it necessitates careful design of reagents and screening of clonal cell lines to avoid disruptive editing at the non-tagged allele. A recent development of split fluorescent proteins has simplified the generation of fluorescent fusions, since only a minimal tag is required for genomic knock-in\textsuperscript{4-6}. Nevertheless, these endogenous tagging methods still require individual HDR donors. Several approaches to develop
generic exon-tagging methods have been demonstrated\textsuperscript{7,8}, but because these require precise tagging at the coding sequence, they are limited in design flexibility and are prone to disruptive mutations at the non-tagged allele, as well as to indels within the tagged allele that can lead to frameshifts. Derivative strategies have been developed to increase the efficiency of homology independent repair-dependent tagging methods but at the cost of no longer utilizing a generic donor\textsuperscript{9}.

An alternative approach for generating endogenous fusions is by random integration of synthetic exons delivered by transposons or retroviral particles\textsuperscript{10}. This approach, known as “protein trapping” or “CD-tagging”\textsuperscript{11}, is not restricted to small donors and has been used in both model organisms\textsuperscript{12-14} and mammalian cells\textsuperscript{15,16}. While protein trapping is inexpensive and scalable, the random nature of tag integration precludes its use for the generation of curated libraries of fusion cell lines.

Here, we combine homology independent repair-based editing\textsuperscript{9} with the use of a generic synthetic exon donor containing a fluorescence tag to perform targeted protein trapping at intronic locations (Figure 1A). This approach is efficient, easy to implement, and does not limit the size of the donor. Furthermore, in contrast to generic exon tagging, generic intron tagging benefits from increased flexibility of the donor design enabled by the splice acceptor and donor sites: any incorporated vector sequence external to those sites has no effect on the coding sequence of the tagged protein. Generic intron tagging also uniquely tolerates mutations in the non-tagged allele, as those are intronic and typically non-disruptive, as well as indels that flank the inserted donor as a result of editing that could lead to frameshifts in an exonic setting. Because the donor is generic, the generation of additional fusion cell lines only requires the cloning of additional intron-targeting sgRNAs.

Initially, we designed a plasmid donor that contained the mNG2\textsubscript{11} tag, part of a recently-published split fluorophore system\textsuperscript{4}, flanked by linker sequences and splice acceptor (SA) and donor (SD) sites (Supplementary Data 1). We embedded this sequence between two identical sgRNA target sites chosen to have minimal off-target activity in the human genome, such that cutting of the plasmid in cells generates a linear DNA donor molecule. Plasmids containing SpCas9, sgRNAs against the donor plasmid, the donor plasmid itself, and intron-targeting sgRNAs were transfected into 293 cells already stably expressing mNG2\textsubscript{1-10}. Proteins successfully tagged with mNG2\textsubscript{11} emit a fluorescence signal upon binding of mNG2\textsubscript{1-10}. Using this approach, we were able to tag four tested genes with well-established localization patterns, CANX, CBX1, VIM and ACTB, at a frequency that enabled easy isolation of both clonal and polyclonal tagged populations of genes (Figure 1B). To test that our tagging approach was indeed mediated
by double-strand breaks in both the genomic sequence and the donor plasmid, we removed each individual component of the transfection mix and found that efficient tagging indeed required all components (Figure 1C). We then tested the feasibility of integrating larger donors by replacing the mNG211 epitope with a full mClover3 fluorescence protein (FP) and found comparable integration efficiencies (Figure 1D). Interestingly, integrating a full-length FP in ACTB resulted in a lower expression level and a diffuse localization pattern, consistent with the production of non-functional protein (right most panels in Figure 1B and 1D). Lastly, to verify that the activity we observed was not specific to 293 cells, we also successfully tagged HeLa cells (Figure 1E).

Unsuccessful tagging can be a result of inefficient sgRNA cutting, low donor integration, inefficient splicing, or a fusion location that disrupts protein expression. To start investigating these alternatives, we chose two genes, ACTB and CANX, and designed nine sgRNAs for each that spanned three introns. We then measured tagging efficiency and the expression levels of tagged cells for each of these locations (Figure 2A,B). We found that both tagging efficiency and expression levels roughly correlated within the same introns, indicating that the location of the fusion is a more critical parameter than the choice of the sgRNA within an intron. We further validated this by testing for donor integration in the total transfected cell populations using pairs of primers that were designed to amplify the genomic DNA to donor junction on both sides of the donor (Figure 2C, diagram). We were able to identify integration for all sites (Figure 2C), with no discernable directional preference (Figure 2C, lower panel). Tandem insertions were also prevalent as evidenced by the upper bands that correspond to twice and sometimes three times the expected molecular weight. We also used Sanger sequencing on some of the amplified junctions and found accurate integration sometimes flanked by junction indels (Figure 2D), further emphasizing the advantages of targeting introns using such an approach.

To summarize, we show that generating endogenous fusions by generic intron tagging is efficient and easy to implement. As editing is done in introns, this approach tolerates indels both in the untagged allele and in the sequences that flank the donor integration site. We show that testing a small number of sgRNAs, spanning multiple introns, is sufficient to identify a successful tagging site, and since these do not require a loci-specific donor, costs are minimal. While more information is needed to better understand how to integrate tags within proteins in the least disruptive way, this approach simplifies the generation of knock-in cell lines and makes scalable gene tagging highly accessible.
Figure 1: Homology-independent generic intron tagging enables efficient and easy generation of endogenous fusions. (A) Illustration of the tagging approach: Double-strand breaks are generated in the intron and donor resulting in the addition of a synthetic intron and fusion of the tag to the coding sequence. (B) Using a small donor composed of the mNG2,11 epitope flanked by splice acceptor and donor sites results in efficient tagging of CANX, CBX1, VIM, and ACTB at the indicated introns (all sgRNA target 1), as observed by flow cytometry (upper panels) and by confocal microscopy (lower panels). Percentages in the dot plots represent the green population as a subset of the total. (C) All transfection mix components are required for tagging of CANX intron 14, sgRNA target 1 (14.1), and of CBX1 intron 3, targets 1 and 2 (3.1 and 3.2). The table indicates which component was removed and bar plots represents the relative percentage of fluorescence-positive cells compared to the full mix. (D) Tagging using a full-length mClover3 fluorophore as a donor. Note that a difference in localization is observed for ACTB between the small and large tag. (E) Tagging of CANX and CBX1 in HeLa cells. All images are maximum projections of z-stacks and scale bars correspond to 10 µm.

Figure 2: Successful tagging is mostly determined by the choice of intron. (A) Tagging with mNG2,11 across introns in ACTB and CANX. Bar plots represent the percent of fluorescence-positive cells for each sgRNA position. (B) Expression mean and standard error for positive cells in each location. Sample sizes are proportional to the bar plots in part A. (C) Gel image showing the amplification of donor to genomic DNA junctions, as illustrated in the right-hand diagrams. Expected band sizes for insertion of a single copy of donor are circled in green. In the diagrams, black arrows represent primer sites for amplification and red arrows represent primer sites for sequencing in part D. The last lane corresponds to a PCR reaction with primers for CANX intron 14, target 1, but without a template. (D) Sanger sequencing of donor to genomic DNA junctions shows de-dephasing at the donor and genomic DNA junction which indicates indels at the integration site.

Methods

Cloning
The mNG2\textsubscript{11} donor tag\textsuperscript{4} flanked by flexible 15 amino acid linkers was synthesized as two complementary oligos from IDT and annealed. This template was amplified by primers to add splice donor and acceptor sites, sgRNA target sequences external to the splice sites, and 25 nucleotide overhangs into the pMC.BESPX-MCS2 parental vector (System Biosciences). pMC.BESPX-MCS2 was digested with EcoRI and Apal, and combined with the mNG2\textsubscript{11} amplicon by Gibson assembly (NEB), generating the pMC-mNG2\textsubscript{11} donor plasmid (Supplementary Data 1). The pMC-mClover3 donor plasmid was generated by replacing the mNG2\textsubscript{11} sequence from the pMC-mNG2\textsubscript{11} plasmid with the sequence of mClover3 (Addgene #74257) by Gibson assembly (Supplementary Data 1).

To generate HEK293 cells stably expressing mNG2\textsubscript{1-10}, mNG2\textsubscript{1-10}\textsuperscript{4} fused to the self-cleaving 2A peptide and tdTomato (Addgene #37347) was cloned into the lenti dCAS-VP64_Blast (Addgene #61425) backbone in place of dCas9-VP64 by 3-piece Gibson assembly.

goRNA-expressing plasmids (Supplementary Table 1) were generated by digesting a lentiGuide-Puro plasmid (Addgene #52963) with Esp31 and ligating an annealed sgRNA oligo duplex as described previously\textsuperscript{3}.

**Cell culture and transfection**

Transfection experiments were carried out in HEK293 (ATCC CRL-1573) and HeLa cells (ATCC CCL-2). The HEK293 cells were generated to constitutively express mNG2\textsubscript{1-10} and tdTomato from a stably integrated lentiviral cassette. Individual clones were sorted based on the tdTomato signal and a line with stable expression over time was selected for experiments.

Cells were cultured in DMEM (Thermo Fischer Scientific) + 10% fetal bovine serum (FBS; VWR) + antibiotic-antimycotic (Thermo Fisher Scientific). 4 x 10\textsuperscript{6} cells were plated across a 12-well plate the day prior to transfection. The donor plasmid was delivered at 5x the molar ratio of lentiCas9-Blast plasmid (Addgene #53962) and the two lentiGuide-Puro plasmids (Addgene #52963) encoding (1) the donor-cutting sgRNA and (2) the genomic locus-targeting sgRNA (Supplementary Table 1). In total, ~1.4 µg of DNA were delivered to each well in 100 µL optimem (Thermo Fischer Scientific) with 4.3 µL of 1g/L PEI (Polysciences, cat. #24765). The following day, each well was expanded into a 10 cm dish. After six days, cells were harvested, analyzed by flow cytometry, and subjected to genomic DNA extraction.
Flow cytometry and cell sorting
Cultured cells were trypsinized, resuspended in DMEM to \(~1 \times 10^6\) cells/ml, and filtered through a cell strainer. Cellular fluorescence was measured on a BD FACSArria Fusion (BD Biosciences). Polyclonal fluorescent cell populations were acquired by isolating 1000 cells by sorting. Data were analyzed using Flowing Software 2 ver. 2.5.1 (http://flowingsoftware.btk.fi/index.php).

Confocal microscopy and image processing
For imaging experiments cells were grown on coverslips and directly fixed in 4% formaldehyde (Electron Microscopy Sciences) in PBS (Thermo Fischer Scientific). Fixed cells were washed in PBS and coverslips were mounted on microscopy slides in Vectashield mounting medium (Vector Laboratories). Images were acquired on a Leica TCS SP8 confocal microscope. Z-stacks (0.6 µm slices) spanning the entire volume of the cells were recorded with oil-immersion 63× Plan-Apochromat lenses, 1.4 NA. Images were processed using Fiji\textsuperscript{17}.

Target analysis
Roughly 2 to 3 \(\times 10^6\) cells were harvested for genomic DNA extraction in 100 µL of QuickExtract (Epicentre) according to manufacturer’s protocol. Amplification of edited genomic regions was performed with the EmeraldAmp MAX PCR Master Mix (Takara Bio USA). Primers were designed using the default parameters of primer3 (http://primer3.ut.ee/) to produce amplicons 250-300 nucleotides in length at the 5’ and 3’ junctions of each targeted site. Amplification reactions included a genomic primer upstream of the target integration site paired with a reverse primer hybridizing to the 3’ end of the tag, or a genomic primer downstream of the target integration site with a forward primer hybridizing to the 5’ end of the tag. The amplicons were imaged and extracted from a 2% agarose gel using the Monarch Gel Extraction kit (New England Biolabs), and analyzed by Sanger sequencing (GENEWIZ) using the tag-hybridizing primers from the amplification reaction.

Bibliography
1 Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science 346, 1258096, doi:10.1126/science.1258096 (2014).
2 Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. Cell 157, 1262-1278, doi:10.1016/j.cell.2014.05.010 (2014).
3 Ran, F. A. et al. Genome engineering using the CRISPR-Cas9 system. Nat Protoc 8, 2281-2308, doi:10.1038/nprot.2013.143 (2013).
4 Feng, S. et al. Improved split fluorescent proteins for endogenous protein labeling. *Nat Commun* **8**, 370, doi:10.1038/s41467-017-00494-8 (2017).
5 Kamiyama, D. et al. Versatile protein tagging in cells with split fluorescent protein. *Nat Commun* **7**, 11046, doi:10.1038/ncomms11046 (2016).
6 Leonetti, M. D., Sekine, S., Kamiyama, D., Weissman, J. S. & Huang, B. A scalable strategy for high-throughput GFP tagging of endogenous human proteins. *Proc Natl Acad Sci USA* **113**, E3501-3508, doi:10.1073/pnas.1606731113 (2016).
7 Lackner, D. H. et al. A generic strategy for CRISPR-Cas9-mediated gene tagging. *Nat Commun* **6**, 10237, doi:10.1038/ncomms10237 (2015).
8 Schmid-Burgk, J. L., Honing, K., Ebert, T. S. & Hornung, V. CRISPant allows modular base-specific gene tagging using a ligase-4-dependent mechanism. *Nat Commun* **7**, 12338, doi:10.1038/ncomms12338 (2016).
9 Suzuki, K. et al. In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. *Nature* **540**, 144-149, doi:10.1038/nature20565 (2016).
10 Trinh le, A. & Fraser, S. E. Enhancer and gene traps for molecular imaging and genetic analysis in zebrafish. *Dev Growth Differ* **55**, 434-445, doi:10.1111/dgd.12055 (2013).
11 Jarvik, J. W., Adler, S. A., Telmer, C. A., Subramaniam, V. & Lopez, A. J. CD-tagging: a new approach to gene and protein discovery and analysis. *Biotechniques* **20**, 896-904 (1996).
12 Buszczak, M. et al. The carnegie protein trap library: a versatile tool for Drosophila developmental studies. *Genetics* **175**, 1505-1531, doi:10.1534/genetics.106.065961 (2007).
13 Clyne, P. J., Brotman, J. S., Sweeney, S. T. & Davis, G. Green fluorescent protein tagging Drosophila proteins at their native genomic loci with small P elements. *Genetics* **165**, 1433-1441 (2003).
14 Trinh le, A. et al. A versatile gene trap to visualize and interrogate the function of the vertebrate proteome. *Genes Dev* **25**, 2306-2320, doi:10.1101/gad.174037.111 (2011).
15 Cohen, A. A. et al. Dynamic proteomics of individual cancer cells in response to a drug. *Science* **322**, 1511-1516, doi:10.1126/science.1160165 (2008).
16 Sigal, A. et al. Generation of a fluorescently labeled endogenous protein library in living human cells. *Nat Protoc* **2**, 1515-1527, doi:10.1038/nprot.2007.197 (2007).
17 Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682, doi:10.1038/nmeth.2019 (2012).
Figure 1

A

B

C

D

E
Figure 2