bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

A strategy for genome-wide seamless tagging of human protein-coding genes

Authors: Karimi, Minoo[1], Saber, Masih M.[1], Stifani, Nicolas[1], Gauthier, Louis[1], Serohijos, Adrian W.[1*] & Michnick, Stephen W.[1*]

[1]Département de Biochimie, Université de Montréal; C.P. 6128, Succursale centre-ville, Montréal, Québec H3C 3J7, Canada.

* Corresponding authors. Email: stephen.michnick@umontreal.ca and adrian.serohijos@umontreal.ca

Abstract

Comprehensive and systematic proteome-wide experiments require financial and technical resources unavailable to most researchers. Here we describe a scalable CRISPR/Cas9-non-homologous end joining (NHEJ) based method for Pooled **R**ecombinant **I**ntegration of **S**eamless **M**arkers (**PRISM**) **into protein coding genes. We created two gRNA libraries for 5'- and 3'-tagging of 18,804 human protein-coding genes. Selection for in-frame integration of the donor cassette can be guaranteed by fusing it to an antibiotic resistance enzyme (ARE) and P2A self-cleaving peptide, resulting in tagged proteins and free ARE to select for clones with inserted donor cassettes. We achieved a library integration rate of 19.75% and tagging of ~80% for genes expressed in Hek293T cells and notably, 89.7% of essential genes, with donor DNA. Our strategy is scalable, specific, and selective, paving the way for genome-scale construction of human cell lines tagged with different types of reporter genes for protein functional characterization.**

Significance statement:

Here we report the first practical method to achieve near complete 5'- or 3'-end integration of reporter protein-coding sequences to express seamless fusions of human protein and reporter proteins that we call **P**ooled **R**ecombinant **I**ntegration of **S**eamless **M**arkers (**PRISM**). PRISM is independent of homology templates, yet it works with high efficiency and precision. Using gRNA-directed CRISPR/Cas9 genome editing and mini-plasmid donor cassettes we were able to tag about 80% of the protein-coding sequences of human genes with a green florescent protein at an

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

integration efficiency of about 20% and most notably, of 90% of essential genes in the human cell line HEK 293T. This is important because the characterization of gene functions in a particular cell type usually begins with essential genes.

Keywords

proteome-wide , CRISPR/Cas9-non-homologous end joining (NHEJ), gRNA libraries, in-frame integration, antibiotic resistance enzyme (ARE), self-cleaving peptide, tagged proteins, essential genes.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Main Text

CRISPR-Cas9 gene editing has revolutionized modern genetics, allowing researchers to precisely knock out genes and regulate their expression at the transcriptional level (1-5). This technology has democratized genetic research, making gene ablation and epistatic analyses widely accessible across laboratories. However, while CRISPR-based gene editing is now routine, large-scale protein tagging remains a challenge. A streamlined approach to tagging proteins would facilitate in vivo tracking of protein localization, abundance, and interactions, greatly enhancing functional genomics (6-8). Currently, efforts to create cell libraries with tagged proteins are limited by technical hurdles. Most existing methods depend on homology-directed repair (HDR), which has low efficiency and requires time-consuming construction of donor DNA with precise homology arms for each of the ~20,000 human protein-coding genes (9). As a result, protein-tagging strategies have been applied to only a limited number of genes and cell lines, restricting their broader adoption.

Despite these limitations, advances have demonstrated the feasibility of large-scale protein tagging. The High-throughput Insertion of Tags Across the Genome (HITAG) method allows efficient tagging of proteins at their C-termini, enabling systematic functional characterization (10). The OpenCell project has further advanced the field by using a split-GFP system, where a small C-terminal GFP fragment is inserted endogenously, complemented by an independently expressed N-terminal fragment, allowing systematic mapping of ~1,300 human protein co-localizations and protein-protein interactions (8). While these techniques represent progress, their application remains constrained by technical limitations, preventing widespread use. Recent studies, including work from the Shalem lab (11, 12) and Reicher *et al.* (13), have explored alternative approaches to improve HDR efficiency, simplify donor DNA design, and enhance scalability. Expanding these strategies is critical for enabling proteome-wide tagging. Developing scalable and accessible protein-tagging technologies could make systematic proteome analysis as routine as CRISPR-based gene editing. These advancements would facilitate large-scale studies in functional genomics, drug discovery, and systems biology, ultimately enhancing our understanding of cellular dynamics and disease pathology.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

An effective strategy for large-scale protein tagging should meet four fundamental criteria: (i) universality, allowing a single donor template to be applicable for tagging diverse genomic loci; (ii) scalability, facilitating the efficient and cost-effective tagging of numerous genes; (iii) specificity, ensuring precise insertion of tags at the target genomic locus without incorporating extraneous sequences such as selection markers or plasmid backbones; and (iv) selective enrichment of successfully modified cells. Here, we introduce a CRISPR-Cas9-based approach for the seamless integration of DNA cassettes into human protein-coding genes at a near-proteome-wide scale. This method, termed Pooled Recombinant Integration of Seamless Markers (PRISM), is broadly applicable to dividing cells, technically accessible to researchers, and satisfies all the specified criteria described above, making it a powerful tool for functional proteomics.

PRISM relies on five key elements: **i**, we use the CRISPR/Cas9 guide RNA directed nuclease SpCas9-NG, which has an expanded repertoire of protospacer-adjacent motif (PAM) specificities (14, 15); **ii**, employing highly efficient non-homologous end joining (NHEJ) repair mechanism; **iii**, a one-cut minicircle DNA plasmid to deliver donor DNA into the cells and cleave the donor plasmid within cells, **iv**, presence of a porcine teschovirus-1 (P2A) self-cleaving peptide (16) fused to an antibiotic resistance gene to select for cells which have the antibiotic resistance gene integrated in-frame and **v**, using standard next-generation sequencing (NGS) of locus specific gRNAs as barcodes to identify and quantify the abundance of tagged genes (Fig. 1) .

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

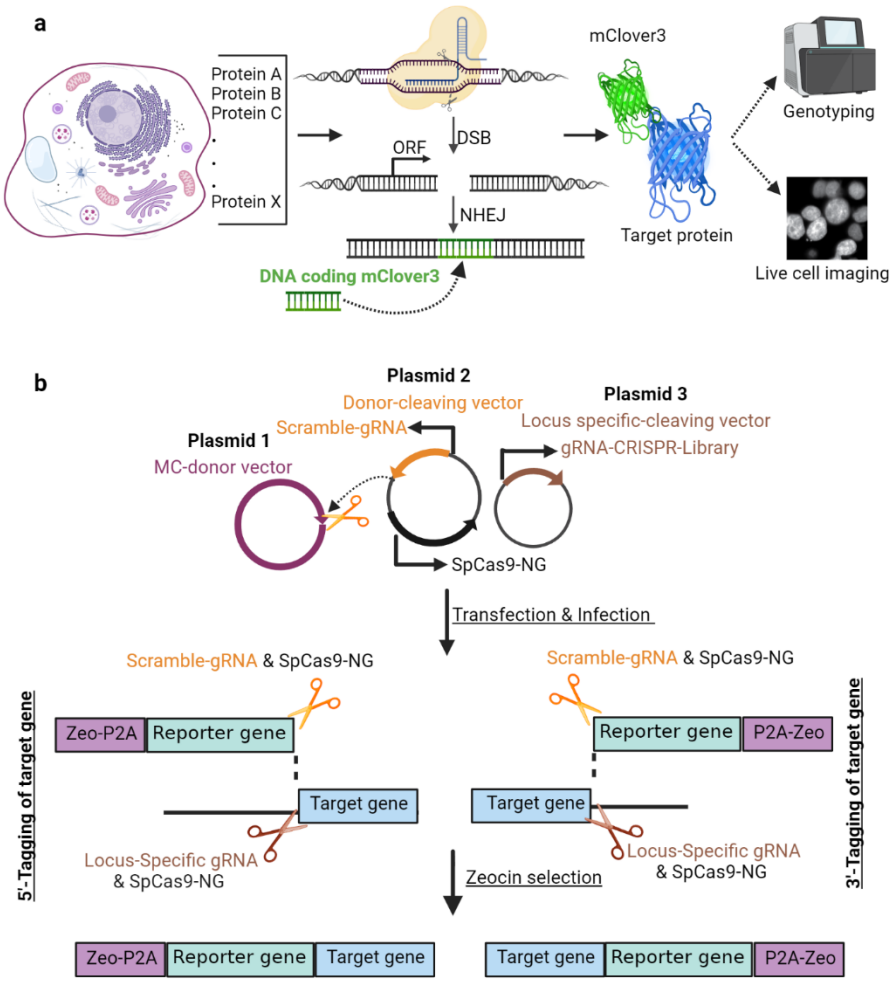


Fig. 1 | Design of CRISPR/Cas9-non-homologous end joining (NHEJ) based method for Pooled Recombinant Integration of Seamless Markers (PRISM). a, Overview of the CRISPR-Cas9 editing strategy to seamlessly integrate DNA cassettes as fusions to human protein-coding genes and ways in which the results have been evaluated. **b**, Schematic of knock-in vectors and overview of gene editing. The system comprises of three plasmids: plasmid 1, a "donor vector" that encodes an antibiotic resistance gene as a selection marker fused to P2A self-cleaving peptide (purple), plasmid 2, a "donor cleaving vector" that carries the Scramble-gRNA (Sc-gRNA) (orange) and Cas9 variant SpCas9-NG (black), and plasmid 3, a "locus-specific vector" that encodes a library of gRNAs targeting 5' and 3' ends of protein-coding sequences of 18,804 human genes that defines the genomic target to be modified by SpCas9-NG (brown).

In general, CRISPR/Cas9-based genome editing begins with the RNA-guided endonuclease Cas9 introducing DNA double-strand breaks at genomic target sites. These double-

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

strand breaks (DSB) then trigger three categories of DNA repair mechanisms: **1**, homology-directed repair (HDR) , **2**, non-homologous end joining (NHEJ) and **3**, microhomology- mediated end joining (MMEJ), each with its own advantages and disadvantages for gene editing (17-21). Although HDR and MMEJ lead to precise repair, their rate is low because they occur only during late S to G2 phases (9, 18, 19, 22). Construction of targeting vectors with the required left and right homology arms (0.5-1 kb) is also cost prohibitive when applied at a large scale. HDR and MMEJ have, however, been successfully applied in several gene knock-in studies, where the low repair rate is offset by increasing the length of homology arms or inhibiting the NHEJ pathway (23-26). Although the NHEJ repair rate is significantly higher than HDR, it results in unproductive out-of-frame or mutant transcripts and thus, NHEJ is known as an error prone repair system (7, 27). To eliminate any errors that could occur by NHEJ, we developed a workaround by designing integrant cassettes consisting of coding sequences for an antibiotic resistance gene (Zeo) fused to a P2A self-cleavage peptide at either its 5'- or 3'-end and a reporter gene, the green fluorescent protein variant mClover3 (28). P2A self cleaves at a rate of ~96% resulting in protein-integrant fusions that are dissociated from the antibiotic resistance enzyme (16). Additionally, using antibiotic selection, only cells in which the antibiotic resistance enzyme was integrated in-frame survive. Although the design of donor cassette guarantees in-frame integration for 3' tagging, in-frame 5' integration is not guaranteed by this strategy, but secondary readouts, such as visualization of fused fluorescent protein, provides a secondary selection assuring that only in-frame fusions are observed (Fig. 1b).

Although NHEJ could still cause in-frame codon deletions, it is unlikely to alter the stability or function of proteins than the fusions themselves. Also, N- and C-termini of most proteins are intrinsically disordered and are thus unlikely to contribute to folding, stability or function (29). Furthermore, we perform 5' and 3' integrations so that either N- or C-terminal fusion proteins are created, providing an extra selection for fusions that do not affect the expression, stability, or function of proteins. Finally, we used single-cut minicircle plasmid (MC) to deliver donor DNA because of their improved integration efficiency and has low toxicity due to its small size and increased lifetime of donor DNA in the cells (30, 31). Furthermore, MCs lack bacterial backbone sequences, eliminating undesired potential integration of backbone plasmid. Overall, PRISM provides a robust means to integrate any large donor DNA cassettes into protein-coding

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

genes at near genome-wide scale with product proteins seamlessly fused to a reporter protein of interest.

Results

Design of seamless DNA cassette integration using NHEJ-based CRISPR-Cas9

We have integrated cassettes corresponding to antibiotic-P2A-mclover3 for functional tagging at N-terminus and mclover3-P2A-antibiotic for tagging at C-terminus. To apply the NHEJ method, we used three individual plasmids (Fig. 1b). Plasmid 1 is a minicircle donor plasmid (MC-donor) harboring the Zeocin resistance gene Zeo, in two orientations: (Zeo)-P2A for 5' tagging and P2A-(Zeo) for 3' tagging. Plasmid 1 also contains a cleavage site complementary to a 20bp gRNA named 'Scramble gRNA' (Sc-gRNA) (The Scramble gRNA was originally described by Suzuki et al. (30) and was identified as a single guide RNA (sgRNA) that does not target any specific gene within the human genome) followed by the PMA sequence TGG (Fig. 1b, Extended Data Fig. 1). Plasmid 2 is a CRISPR/Cas9 expression plasmid that harbors the coding sequence for SpCas9-NG under control of a CMV promoter (15). SpCas9-NG has an expanded repertoire of PAM specificities which enables the design of more than one gRNA per targeted integration locus of each gene to introduce in-frame DSB. Plasmid 2 also contains the Sc-gRNA that introduces a cut into the donor DNA (Fig. 2a and Extended Data Fig. 2a). Sc-gRNA is particularly chosen because it is not present in any protein-coding gene in the human genome, which eliminates the possibility of off-target activity (30). Plasmid 3 contains the lenti-CRISPR gRNA library that targets all ~20K ORFs of the human genome (Fig. 2b and Extended Data Fig. 2b).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

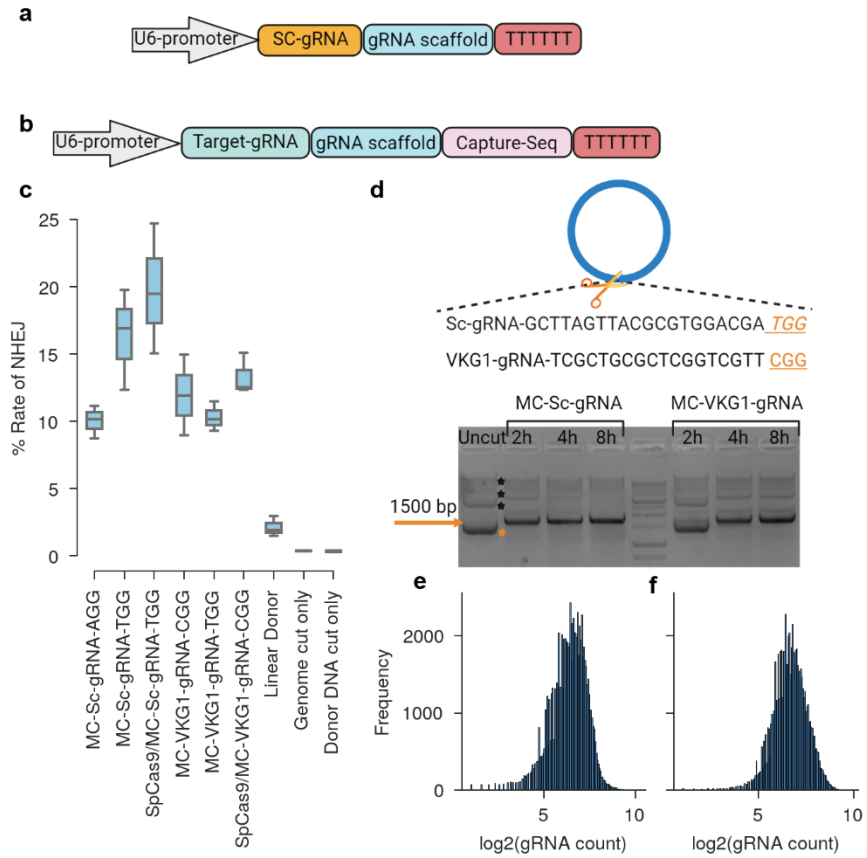


Fig. 2 | Design and optimization of seamless DNA cassette integration into the genome using NHEJ-based CRISPR-Cas9. a, Design of donor-cleaving plasmid. The U6 promoter-based expression cassette followed by the Sc-gRNA (orange), gRNA scaffold (blue) and termination signal sequence (red). **b**, Design of gRNA-CRISPR-library plasmid. The U6 promoter-based expression cassette followed by the library gRNAs targeting 18,804 human genes including their splice variants (cyan), gRNA scaffold (blue), capture sequence (pink) and termination signal sequence (red). **c**, Efficiency of NHEJ-based strategies using two different gRNA cleaving vectors. The efficiency of CRISPR/Cas9-induced NHEJ-mediated DNA integration was tested by targeting five human protein-coding genes to insert four types of minicircle (MC) donor vectors targeted by distinct guide RNAs directed against different PAM sequences, TGG, AGG, or CGG (MC-Sc-gRNA-TGG, MC-Sc-gRNA-AGG, MC-VKG1-gRNA-TGG and MC-VKG1-gRNA-CGG). Results were obtained for the five human genes from three replicates and represented as average ± standard error of the mean. We observed higher knock-in efficiency with Sc-gRNA donor using TGG as PAM (MC-Sc-gRNA-TGG) in comparison to VKG1 donors (MC-VKG1-gRNA-TGG and MC-VKG1-gRNA-CGG). Consistent with previous observations, PCR-amplified donor DNA

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

(linear DNA) resulted in little to no knock-in events (32). We also obtained higher integration efficiency in cells that were transfected with SpCas9-NG plasmid prior to introducing donor vectors and lentiCRISPR-22-gRNAs lentivirus-infected plasmids (SpCAS9-NG/MC-Sc-gRNA-TGG and SpCAS9-NG/MC-VKG1-gRNA-CGG). **d**, *In vitro* Cas9 cleavage Assay to evaluate the efficiency of our strategies for introducing double-strand breaks (DBS) into the donor vectors harboring Sc-gRNA or VKG1-gRNA target sites. The uncut DNA (lane 2) shows supercoiled conformation of minicircle DNA marked with asterisk in orange (*) and 3 possible conformations for parental plasmid marked with asterisks in black (*) (supercoiled, relaxed and nicked). Successful DSB into the donor vectors results in a band of 1600 bp in length. These results show that Cas9 efficiently introduces DSB into Sc-gRNA site within 2 hours of incubation. **e** and **f**, CRISPR library prepared for 5' and 3' tagging respectively. NGS data illustrates the close to equal distribution of gRNAs within the library.

Optimization of seamless DNA cassette integration using NHEJ-based CRISPR-Cas9

To test and optimize our experimental design to endogenously tag protein-coding genes, we selected a set of five human genes including RPS6KB1, AKT1, PKD1, RPS6 and PPP2CA, at both 5'- and 3'-ends, chosen because we previously tagged these genes without detrimental consequences to function (33). Twenty-two gRNAs were designed for the five human genes (Supplementary Table 1) and CRISPR/Cas9- NHEJ-mediated DNA integration was performed by infecting cells with a pool of twenty-two-locus-specific gRNA-expression vectors and at the same time, co-transfecting them with pX330-SpCas9-NG expressing the SpCas9-NG endonuclease and either VKG1-gRNA or Sc-gRNA guides and vector harboring one of four of minicircular (MC) Zeo donor vectors with different PAM sequences TGG, AGG, or CGG (Sc-gRNA-TGG, Sc-gRNA-AGG, VKG1-gRNA-TGG or VKG1-gRNA-CGG), targeted by guides Sc-gRNA or VKG1-gRNA. We tested SpCAS9-NG activity at the VKG1-gRNA or and Sc-gRNA targeted sites. In the original studies, Sc-gRNA and VKG1-gRNA were linearized by wild-type Cas9 within the cells and their PAMs were AGG and CGG, respectively. However, since we used SpCas9-NG instead of wild-type Cas9, we tested TGG as a PAM for these two gRNAs. This is because the SpCas9-NG variant we used has been reported to exhibit the highest cleaving efficiency at TGG (30, 32, 34). After infection and co-transfection, we selected for cells successfully infected with

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

the pool of twenty-two-locus-specific gRNA vectors by treating them with 4 µg/ml of Blasticidin 48 hours after infection (Extended Data Fig. 3a). Then, five days after infection/co-transfection, resistant cells were treated with Zeocin (100 µg/mL) for 14 days (Extended Data Fig. 3b). As negative controls, we also measured the integration efficiency in cells infected/transfected with the locus-specific gRNA vectors and pX330-SpCas9-NG without donor cassette integration and cells co-transfected with pX330-SpCas9-NG harboring either VKG1-gRNA or Sc-gRNA and minicircular donor plasmids, resulting in cleavage of donor plasmids but not specific gene loci. The absolute knock-in efficiencies *via* NHEJ-mediated targeted integration were measured by the ratio of live cells treated with *versus* without Zeocin (Fig. 2c). No knock-in was observed in the two negative controls. In the case of Sc-gRNA, we observed higher knock-in efficiency with MC-Sc-gRNA-TGG (16.34 ± 3.73%) compared to MC-Sc-gRNA-AGG (10.01 ± 1.22%). For VKG1-gRNA, we observed higher knock-in efficiency using MC-VKG1-gRNA-CGG (11.92 ± 3%) than MC-VKG1-gRNA-TGG (10.29 ± 1.1%). Notably, we observed higher knock-in efficiency with Sc-gRNA donors than with VKG1-gRNA. Our data also indicated that the knock-in efficiency was higher using TGG as PAM for Sc-gRNA. Consistent with previous observations, PCR amplified donor DNA resulted in little spurious knock-in events (2.12 ± 0.74%) (32). As successful gene editing requires a minimum threshold expression level of Cas9, we compared the knock-in efficiency between cells that were infected and co-transfected with the three types of editing and donor plasmids at the same time with cells that were first transfected with pX330-SpCas9-NG for 18 hours prior to infection/transfection with the twenty-two-gRNA library and donor plasmids. We observed higher knock-in efficiency when the cells were transfected with SpCas9-NG plasmid prior to introducing target specific gRNAs and donor DNA (SpCas9/MC-Sc-gRNA-TGG: 19.75 ± 4.48 %) (Fig. 2c). We confirmed the efficiencies of MC-Sc-gRNA-TGG and MC-VKG1-gRNA-CGG-directed cleavage in cells with a biochemical cleavage assay. With purified donor plasmids and Cas9 endonuclease, Cas9 cleaved MC-Sc-gRNA-TGG with 90% efficiency compared to MC-VKG1-gRNA-CGG donor vectors within 2 hours of incubation, consistent with our results in cells (Fig. 2d). Collectively, we concluded that using MC-gRNA-TGG as the donor DNA and transfecting cells with pX330-SpCas9-NG prior to infection/transfection with the gRNA library and donor plasmids with knock in efficiency of 19.75 % can be applied to the library-level integration.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Genome-scale lentiviral gRNA library construction

To test whether our experimental design could be scaled to the library-level generation of endogenously tagged human cell lines, we targeted the entire human protein coding genes for donor cassette integration. We used the following criteria in designing a genome-wide gRNA library: 1, gRNAs only introduce double strand breaks in frame; 2, minimal off-target activity; 3, maximal on-target activity; 4, avoidance of homopolymer stretches (e.g., AAAA, GGGG) (35); and 5, GC content between 20 to 80%. This resulted in two gRNA libraries consisting of 73,909 and 85,394 unique guides designed for 5'- and 3'-end tagging, respectively (Supplementary Tables 2 and 3). These libraries contain a minimum of 1 and maximum 8 gRNAs per splice variant of human genes (Extended Data Fig. 4a, b). gRNAs were synthesized as 79-mer oligonucleotides on microarray chips and amplified by PCR as a pool. The PCR products were cloned into a modified version of lentiviral vector 'lentiCRISPR-V2' to yield at least 2000-fold representation per gRNA in each library. We evaluated the diversity and coverage of the library using next-generation sequencing (NGS) (Supplementary Tables 4 and 5). This analysis showed 99.8% and 99.9% coverage of the designed sgRNAs for each N and C library. In addition, A skew ratio below 5 indicated uniform representation, minimizing bias and optimizing screening efficiency. Guide representation ranged from 5% to 95%, ensuring proportional distribution across the library, and maximum screening efficiency and identification of a hit (Fig. 2e, f).

In vivo genome-wide editing and fusion of reporter sequences to protein-coding genes

To test PRISM across the 18,804 protein coding genes, we tested for integration of the coding sequence for the green fluorescent protein variant mClover3 fused to Zeo in HEK293 cell line. Similar to our small-scale test above, the donor DNA had two orientations, P2A self-cleaving peptide preceded or followed by a ten amino acid linker $(Glycine_4Serine)_2$ (linker-mClover3-P2A-Zeo and Zeo-P2A-mClover3-Linker) (Extended Data Fig. 5). With these genes expressed in-frame with the protein-coding sequences, we could select for integration of the donor cassette for Zeo and for functional integration by visualization of mClover3 fluorescence. HEK 293T cells were transfected with pX330-SpCas9-NG/Sc-gRNA and 18 hours later, cells were infected with 5' or 3' gRNA libraries at a multiplicity of infection (MOI) 0.3 and transfected with minicircle plasmid

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

encoding the donor cassette. To select for successful infection of library gRNAs, cells were treated with 4 µg/ml Blasticidin 48 hours post transduction (Extended Data Fig. 3a). Five days later, to select for successful integration of donor cassette, cells were treated with 100 µg/mL Zeocin (Extended Data Fig. 3b) until all the non-transduced cells in the negative control died out. Afterwards, cells were kept in culture for 21 days, passaging every 4 days. To avoid bottlenecking, we aimed to keep the total cell count above 3.7×10^7 and 4.3×10^7 for 5'- and 3'-end tagging, respectively, corresponding to a 500-fold coverage of the library diversity cells. Successful integration was evaluated by imaging cells for mClover3 fluorescence. We observed fluorescence across mixed cultures of selected cells with heterogeneous expression levels inferred from the intensity of fluorescence and unique subcellular localizations suggesting successful functional tagging of a wide variety of genes (Fig. 3a). After the screen was completed, deep sequencing was performed to identify the coding sequences of gRNAs present in the pool of genomic DNA and to measure changes in the gRNA distribution that resulted from the applied antibiotic selection pressure. To do so, we amplified a region of the CRISPR-gRNA-library corresponding to the locus-specific-gRNA/gRNA-scaffold and performed deep sequencing using Nextseq Illumina. Briefly, genomic DNA was extracted from one sample of the cells with successful integration for each site (5' or 3') and subjected to one round of PCR (Extended Data Fig. 6) followed by Illumina next-generation sequencing and aimed for 130 million total reads. In the next step, the NGS data were analyzed with a custom Python script to map reads to the entire designed oligo pool. Based on the number of locus specific gRNAs that were obtained after Zeocin selection, 11,922 genes were tagged in total with 3,022 tagged at both 5'- and 3'-ends, 5,385 uniquely tagged at 5' and 3,515 tagged only at 3' ends of protein-coding sequences (Fig. 3b, and Supplementary Tables 6 and 7). Considering that 15,015 genes are reported to be expressed in HEK cells, we achieved approximately 79% tagging efficiency in HEK 293 cells. The majority of the recovered guide RNA (gRNA) sequences corresponded to those resulting in the loss of zero or only one codon (Extended Data Fig. 7), implying that integrations were preferentially selected for the preservation of codons with minimal disruption to the resulting protein product. We hypothesized that one reason why 9.8% of the proteome could not be tagged is due to their low rate of expression, and consequently, low protein abundance. We used a publicly available ribosome profiling dataset for HEK 293T cells as a benchmark to determine protein expression levels for the successfully and unsuccessfully tagged proteins (36). Ribosome profiling is a sequencing-based method that measures ribosome

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

density on cellular mRNAs, which is indicative of protein synthesis rate. The density of ribosomes for each gene is represented by the reads per kilobase of transcript per million mapped reads (RPKM) value. As protein abundance is closely linked to the rate of its synthesis, RPKM data serves as a reasonable estimate for absolute protein expression levels. Our observations revealed that low protein abundance was a main obstacle to successful selection for cells expressing the tagged proteins (Fig. 3c). Finally, essential genes are often the targets of most scrutiny for functional characterization in a particular cell type but can also be most sensitive to manipulation. PRISM should work as a filter, selecting for N- or C-terminal fusions of proteins to essential gene products that do not result in functional disruption, due to destabilization, degradation or mis-localization of the protein. To assess the capability of PRISM for tagging essential genes, we examined the number of locus-specific gRNAs representing essential genes obtained following Zeocin selection. 680 essential genes were tagged at both 5'- and 3'-ends (N- and C-terminal tagged), 304 uniquely tagged at 5' (N-terminal tagged) and 283 tagged only at 3' ends (C-terminal tagged) of essential protein-coding sequences (Fig. 3d). This amounts to an 89.7% success rate (1,267¸ 1,412). In contrast a systematic C-terminal tagging effort would only achieve 68.2% success (963¸ 1,412). Based on RPKM data, low protein abundance could be a major obstacle to successful tagging of remaining 10.3% of essential genes (Fig. 3e). These results illustrate a key application of PRISM, creating a roadmap for the seamless and systematic integration of tags into protein coding genes including essential genes, to generate arrays of cell lines with specifically labelled proteins. Essentially, PRISM is a screen that allows for selection of the best sgRNAs and preferable N- or C-terminal tagging orientation to minimally disturb the function of a protein. Thus while as described, PRISM is meant to be used as a pooled screening approach, it would also vastly simplify and reduce costs of creating systematically tagged protein genes in any dividing cell line, primary or IPS cells.

We next determined whether PRISM results in correct fusions to protein coding genes by examining the localization of mClover3 at the single cell level. To achieve this, we randomly selected 96 single cells from our mClover3-pooled library, which we expanded and imaged each individually. Out of the 96 cells, 74 cells displayed detectable signal over background (77%). For each of these 74 targets, each cell was analyzed by microscopy, and the target protein was identified through sequencing. Then, we compared our localization annotations with the Human

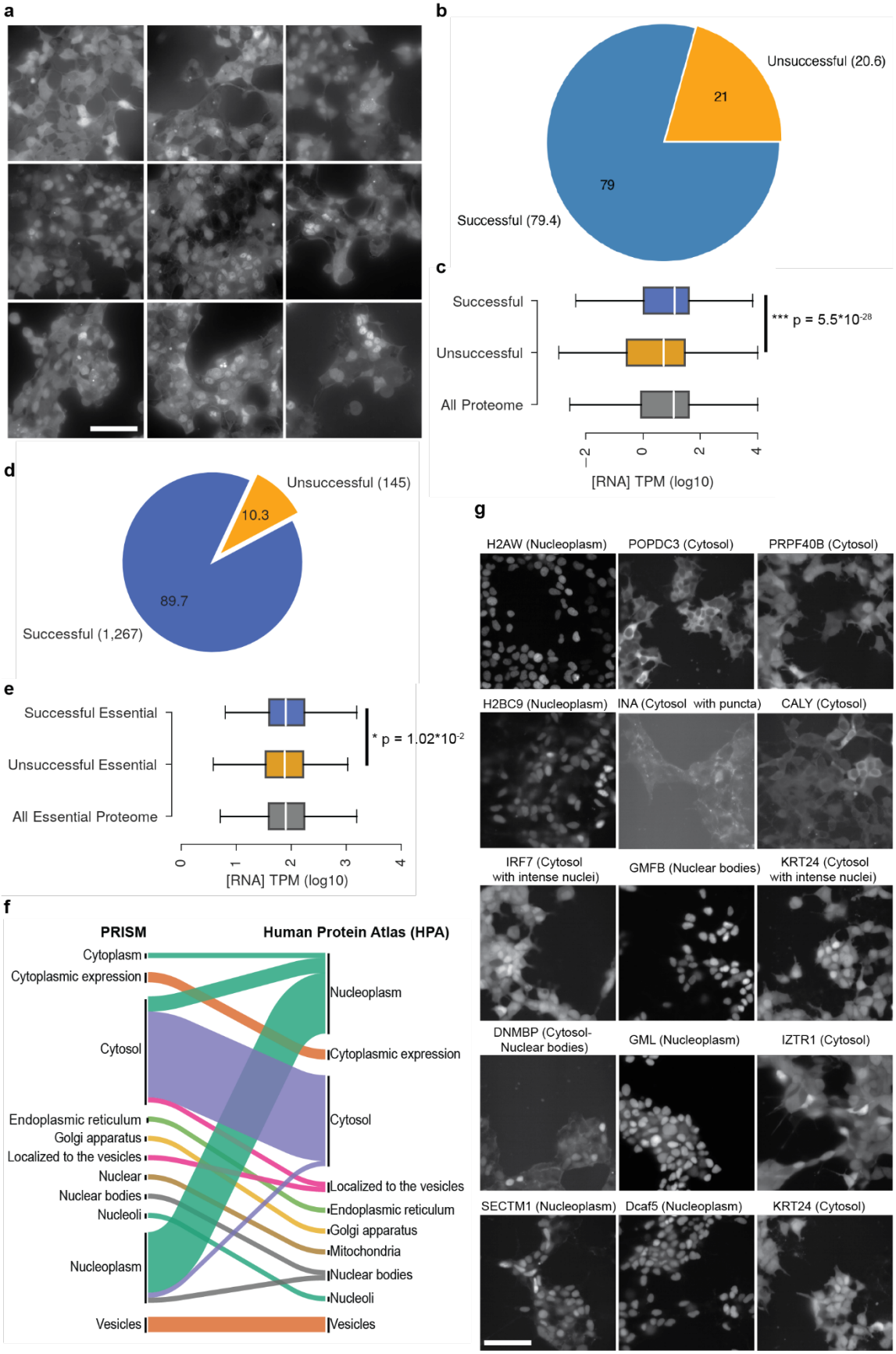
bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Protein Atlas (HPA), which is a reference antibody-based collection of human protein localization (37). The comparison showed a significant agreement between datasets, as 70% of proteins shared the same localization annotation (Fig. 3f). It is important to note, however, that HPA mainly reports on cell lines other than HEK 293T, so we did not expect a perfect overlap because proteins can vary in their localization across related compartments in different cell types. (Fig. 3g and Extended Data Fig.8).

For PRISM to be considered universally applicable, it must work in other cell lines. To test this, we implemented PRISM in four additional cell lines: HCT116, Vero, HeLa, and Hap1. While we initiated CRISPR-NHEJ knock-in with 1000-fold coverage in HEK 293T cells, we opted for a lower 300-fold coverage in the other cell lines as a simplified proof of concept. Our results confirmed successful gene integration across all four cell lines, with knock-in efficiencies of 10.58 % (1,909 genes) in HCT116, 15.45 (2,905 genes) % in Vero, 25% (4,703 genes) in HeLa, and 20.97 % (3,944 genes) in Hap1(Extended Data Fig. 9 and Supplementary Table 8). Although the efficiency varied, two primary factors influenced these results: the inherent transfection efficiency in each cell line and the fold coverage of the CRISPR library used in each experiment. Improving editing efficiency in other cell lines will require increased fold coverage of the CRISPR library and modifications and optimization of our transfection protocol.

While in HEK 293T cells, we were able to perform transfection and lentiviral infection simultaneously without affecting the efficiency of either process, this was not the case for other cell lines. Specifically, in Hap1 and HCT116 cells, a sequential approach—first conducting lentiviral infection, expanding the cells, and subsequently transfecting them with the plasmid encoding Cas9 and donor DNA—resulted in improved knock-in efficiency.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

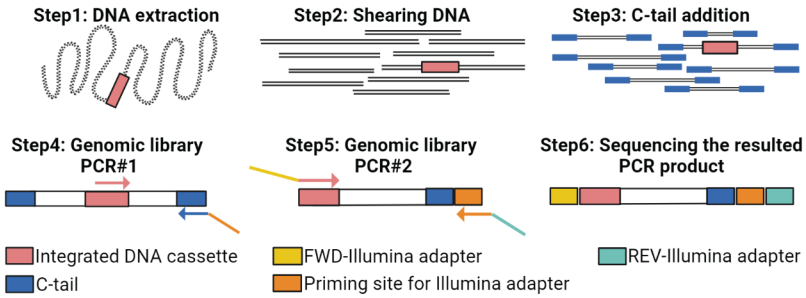
Fig. 3 | Genome-wide CRISPR integration of Reporter protein donor DNA fragments. a, Visualization of mixed pools of cells in which endogenous proteins are fused to mClover3 using our CRISPR/Cas9-NHEJ-based strategy. Cells vary in fluorescence intensity and patterns of expression suggesting tagging of many proteins with varying abundances and different subcellular localizations. Scale bar is 50μm. **b**, NGS results indicate that ~80 % of the human genes were tagged 5' or 3' to protein-coding sequences in the genome of HEK293T cells. **c**, The graph shows the distribution of abundance for all proteins expressed in HEK293T versus successfully or unsuccessfully detected targets; boxes represent 25th, 50th, and75th percentiles, and whiskers represent 1.5 times the interquartile range. Median is indicated by a white line. Outliers are not shown. Statistical significance p-value is 5.5×10^{-28} and is calculated by Student's t-test. Data indicates that low protein abundance posed a significant challenge to successful tagging, as the proteins that were not successfully tagged displayed low or no expression levels in HEK 293T cells. **d**, NGS results indicate that 89.7% of the essential genes were tagged 5' or 3' to protein-coding sequences in the genome. **e**, the distribution of abundance of essential proteins expressed in HEK 293T *versus* successfully or unsuccessfully tagged genes; boxes represent 25th, 50th, and 75th percentiles, and whiskers represent 1.5 times the interquartile range. Median is indicated by a white line. Statistical significance p-value is 1.02×10^{-2} and is calculated by Student's t-test. Data indicates that low protein abundance posed a challenge to successful tagging of essential genes, as the proteins that were not successfully tagged displayed low expression levels in HEK 293T cells. **f**, comparison of annotated protein localizations between PRISM and the Human Protein Atlas (HPA) datasets (37). The diagram features colored bands that correspond to groups of proteins with similar localization annotations between our data set and HPA. The width of each band is proportional to the number of proteins in the group. **g**, imaging analysis of individual successful targets. mClover3 fluorescence intensity and subcellular localization vary widely for each gene.

Direct verification of integration of reporter DNA cassette as fusions to protein-coding sequences

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

We sequenced a portion of the tagged libraries to directly confirm reporter DNA cassette integration by semi-random PCR in which one of the designed primers targets a part of the donor cassette and the second random primer recognizes the adjacent sequences of the protein coding sequence of the gene to which donor cassette is fused (Supplementary Table 9). Samples were sequenced using Illumina NextSeq500 high output 2x75bp Flow Cell (Fig. 4a). The correct paired reads with forward and reverse reads mapped within the defined DNA fragment length on the human genome were then sorted and the read pairs overlapping with genomic locations targeted by gRNAs in the library were retrieved. Our data confirmed in-frame integration of reporter DNA cassettes as fusions to protein-coding sequences of genes (Fig. 4b, c and Supplementary Table 10).

a



b

Genes	Donor DNA	Target gene
EEF2	GGT GGG TCC CCA TCG	ATG GTG AGT AAC CCT GCC
ACTB	GGT GGG TCC CCA TCG	ATG GAT GAT GAT ATC GCC
PLCD1	GGT GGG TCC CCA TCG	ATG CAG TGC CTG GGG ATC
H4-16	GGT GGG TCC CCA TCG	ATG TCT GGG CGA GGT AAA
CCND3	GGT GGG TCC CCA TCG	ATG GAG CTG CTG TGT TGC
TRIM4	GGT GGG TCC CCA TCG	ATG GAA GCT GAG GAC ATC

c

Genes	Target gene	Donor DNA
TUBB4A	GCG GAG GAG GAG GTG GCC	CGA TGG GGC GGT GGC
HMGB3	GAG GAG GAG GAG GAT GAA	CGA TGG GGC GGT GGC
NADK	GAG GAG GAG GAG CGA TGG	CGA TGG GGC GGT GGC
H2BC10	AAG TAC ACC AGC TCC AAG	CGA TGG GGC GGT GGC
H2BU1	AAG TAC ACC AGC TCC AAG	CGA TGG GGC GGT GGC
H2BC18	AAG TAC ACC AGC TCG AAG	CGA TGG GGC GGT GGC

Fig. 4 | Identification of donor DNA insertion site by nested PCR. a, Nested PCR strategy to identify the locations of donor DNA in the genome. First, genomic (gDNA) was extracted from library-containing cells and sheared by sonication into fragments of between 200 and 600 bp. Next, terminal deoxynucleotidyl transferase (TdT) was used to add repeats of CCCCCCCCCCCCCCCC (C-Tail) at the end of fragmented DNA. The C-tailed genomic library was subjected to two rounds of PCR. The first PCR was performed with a forward primer that recognized a part of the donor fragment and the reverse primer was designed to hybridize to the C-Tail (GGGGGGGGGGGGGGGG). Product of the first PCR was subjected to the second round of PCR.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

399 This step enabled the addition of the adapters necessary for sequencing. The resulting PCR
400 products were subjected to PCR clean up followed by Next Generation Sequencing (NGS). **b-c**,
401 Examples of observed protein-coding sequences fused 5' or 3' to donor DNA. Sequences of the
402 donor DNA blue are fused to protein-coding sequences without or with missing protein coding
403 codons (red).
404
405 **Discussion**
406
407 Here, we report a robust approach for the rapid and efficient generation of reporter protein-tagged
408 endogenous human gene sequences. Critical to the efficiency of PRISM was the use of circularized
409 mini plasmids as donor vectors and to cleave them intracellularly cleavage. Plasmids have a longer
410 lifetime than linear DNA, and proteins involved in NHEJ, such as KU70 and KU80, can attach to
411 the 5′- or 3′-ends of donor DNA in response to DSB in cells, which protect donor DNA from
412 degradation and facilitate ligation of the donor to genomic DNA (38). Compared with existing
413 methods, PRISM has the following advantages: First, in contrast to multistep preparation of HDR
414 targeting vectors, we designed donor cassettes such that the donor vectors for all ~20,000 human
415 proteins can be generated rapidly and in large scale without any requirement for target sequence-
416 specific donor vector reconstruction. Seamless HDR tagging of protein genes will remain
417 important for systematic, quantitative spatiotemporal analysis of proteins in living cells, and
418 measuring protein abundances and protein-protein interactions (7, 8). However, PRISM will aid
419 in screening for functional fusions of genes that have proven recalcitrant to tagging with reporters
420 and to obtaining endogenous fusion of many more genes than are currently available and in
421 different cell lines, iPS cells and dividing primary cells. Second, recruitment of SpCas9-NG and
422 minicircle donor plasmid leads to high knock-in efficiency while minimizing off-target cleavage
423 and eliminating undesired integration of backbone plasmid. Third, the P2A-antibiotic donor
424 system in our design provides a simple solution to select only cells with in-frame 3' donor DNA
425 and eliminates the errors introduced by NHEJ while any out of frame 5' integrations will not appear
426 in any secondary selection or measurements performed on cells expressing the fusion proteins.
427 Finally, the utility of endogenously tagged proteins enables the function of a protein to be
428 characterized under the control of native regulators of gene expression and without disturbing
429 endogenous interaction stoichiometry.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

We envision that PRISM could be used to manipulate protein abundances or interactions in a variety of ways, including chemically-induced degradation or proximity, such as auxin-induced single or pairwise protein degradation or mandipropamid-induced forced proximity of protein pairs (39-42). Equally, measuring protein abundances or protein-protein interactions could be achieved with protein-fragment complementation assays, such as those based on dihydrofolate reductase reporter (43, 44) or split fluorescent proteins (8, 45). Each of these methods would provide measurements *via* depletion or enrichment of sgRNA-coding sequences that are integrated into the genome as measured by NGS, for individual or pairs of protein-coding genes, due to fitness effects of the protein manipulation or interactions making these methods fully accessible to most labs. We could also envision direct measurements, for instance, measuring protein abundance by binning single cells expressing proteins fused to fluorescent proteins by fluorescent intensity followed by NGS of the individual bins.

Altogether, PRISM provides scalability, specificity, selectability and paves the way for the genome-scale construction of human cell lines tagged with different functional tags at endogenous loci, with a simple NGS readout of sgRNAs to identify protein genes that are tagged in individual cells. Consequently, PRISM makes whole proteome tagging and functional genomics readily accessible to any researcher aiming to manipulate or measure the abundances or interactions of proteins.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Methods

Plasmid construction

Constructing plasmid harboring SpCAS9-NG and donor cleaving gRNA

The plasmid coding Cas9-NG (34) was purchased from Addgene (PX330- SpCAS9-NG). PX330-SpCAS9-NG has two expression cassettes, a human optimized SpCAS9-NG and a single guide RNA. The guide RNA expression cassette from PX330- SpCAS9-NG plasmid was used to individually clone two donor-cleaving gRNAs named "scramble-gRNA" (Sc-gRNA) GCTTAGTTACGCGTGGACGA (30) and VKG1-gRNA TCGCTGCGCTCGGTCGTT(32). The gRNA sequences were synthesized as single stranded DNA (Supplementary Tabl0e 1) amplified by PCR and cloned 3' to a U6 promoter following the protocol of Ran *et al.* (46). Briefly, gRNA forward and reverse primers (Supplementary Table 11) were annealed to form duplexes in a mixture of equal molar concentrations and volumes in a heating block heated to 100˚C and gradually cooled to room temperature. The duplex was then used as an insert ligated by T4 DNA ligase (New England Biolabs) into the coding sequence of PX330-SpCAS9-NG digested by BbsI (New England Biolabs). The ligation products were transformed into Stbl3-competent cells and the resultant clones were screened by cracking gel for insert sizes and verified by sequencing. The Sc-gRNA and VKG1-gRNA used here do not exist in human, mouse or rat genomes and are reported to successfully guide the introduction of double-strand breaks into a donor DNA flanked with complementary sequences using Wild type Cas9 (30, 34).

Constructing donor vector for 5' and 3' tagging of protein-coding gene sequences

To make donor vectors Zeocin-P2A-mClover3-linker and linker-mClover3-P2A-Zeocin for tagging human genes at respective 5' and 3' ends, the coding sequence of the Zeocin resistance gene Sh ble was PCR amplified from the LentiCRISPR-V2 plasmid (Addgene, Plasmid #52961) , P2A was amplified from LentiCRISPR-V2 plasmid (Addgene, Plasmid #52961) using P2A-FWD (forward) and P2A-REV (reverse) primers (Table S10) and the coding sequence of mClover3 was amplified from pMK290 (mAID-mClover-Hygro) plasmid (Addgene, Plasmid #72828). All the

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

fragments were assembled *in vitro* using Gibson assembly (New England Biolabs) and were flanked with complementary sequences of Sc-gRNA or VKG1-gRNA followed by either AGG or TGG as PAMs for the Sc-gRNA and CGG or TGG for VKG1-gRNA (Supplementary Table 12). The tested and reported PAMs for Sc-gRNA and VKG1-gRNA are AGG and CGG, respectively. Both were tested using wild type cas9. We also tested the TGG as PAM for these two gRNAs because the SpCas9-NG variant we used been reported to have highest cleaving efficiency at TGG sequences (34). In the next step, the fragments were cloned into ApaI and SmaI sites of the minicircle producer plasmid pTubb3-MC (Addgene, Plasmid #87112) using GeneArt seamless cloning (Thermo). In this design, the expression of the Zeocin resistance marker and mClover3 genes are driven by the endogenous promoter of the protein coding genes to which the mClover3-coding oligonucleotide donor is fused.

Construction of Lenti-CRISPR-gRNA library

We used a modified version of LentiCRISPR-V2 (47) plasmid for cloning the gRNA libraries. In this plasmid, we first replaced the sequence coding for Cas9 with that for mCherry using PCR followed by GeneArt seamless cloning (Supplementary Table 12). This plasmid includes an expression cassette for gRNA followed by a constant region named gRNA scaffold (also called trans-activating crRNA (tracrRNA) (Extended Data Fig. 2b). We then added a short sequence named “Capture Sequence” (GCTTTAAGGCCGGTCCTAGCAA) (48) immediately before the gRNA termination signal, which elongates the 3’ end of the gRNA (Extended Data Fig. 2b, Supplementary Table 12). The capture sequence is a 22 bp DNA fragment, which was designed and used in feature barcoding technology to directly capture gRNAs present in a cell inside a gel-bead-in-emulsion (48). Dixit *et al.* showed that this short fragment does not perturb the gRNA scaffold, so we used this sequence to provide a priming site to later amplify the gRNA library from the Lenti-CRISPR plasmid and to help distinguish the donor-cleaving gRNA from the gRNA library in later steps[39]. We did this because both gRNA library and donor cleaving-gRNA have the same promoter and scaffold sequences, thus, the capture sequence provides a priming site for PCR amplification of sgRNA-coding sequences in individual cells for next-generation sequencing to infer the genes and in which 5’ or 3’ orientations the donor DNA has been fused (Extended Data Fig. 2). gRNA scaffold followed by the Capture Sequence was synthesized as a single stranded

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

DNA (IDT) (Supplementary Table 12) and was cloned into the LentiCRISPR-V2 plasmid using GeneArt seamless cloning kit (Thermo Fisher Scientific).

Minicircle DNA vectors preparation

Production of minicircle DNA vectors (MC) were performed in accordance with the manufacturer's protocol (System Biosciences (SBI)) with some modification. Briefly, the final minicircle constructs, pre-minicircle plasmids P2A-Zeocin-MC, Zeocin-P2A-MC and linker-mClover3-P2A-Zeocin-MC with Sc-gRNA targeting sequences were introduced into the ZYCY10P3S2T E. coli minicircle producer competent cells (SBI) and amplified overnight in TB (Terrific Broth) (pH 7.0) (Fisher Scientific). The minicircle production was induced by mixing the overnight TB culture with an equal volume of minicircle induction mix comprising fresh LB and 20% l-arabinose, followed by a 5 h incubation at 30°C with shaking at 250 rpm and 2 h incubation at 37°C with shaking at 250 rpm. Minicircle DNA was then isolated with ZymoPURE II Plasmid Maxiprep Kit (Zymoreserach). Finally, 1.0 µg of each minicircle plasmid (P2A-Zeocin-MC, Zeocin-P2A-MC and linker-mClover3-P2A-Zeocin-MC) was digested with SmaI restriction enzyme (New England Biolabs) and then run on a 1.5% agarose gel to determine plasmid quality (Extended Data Fig. 10).

***In vitro* Cas9 cleavage assay for donor cleaving gRNAs efficiency**

To cleave the donor plasmid, Sc-gRNA and VKG1-gRNA were chosen as the donor-cleaving gRNAs because they do not introduce any double-strand break into the human genome[18,21]. We checked the cleavage efficiency of both Sc-gRNA and VKG1-gRNA toward donor DNA using an *in vitro* Cas9 Cleavage Assay. To prepare each donor-cleaving gRNA for *in vitro* transcription, gRNAs were prepared following methods by Lin *et al.* with some modifications[40]. gRNAs were obtained by *in vitro* transcription of a DNA template of the following sequence: 5′-TAA TAC GAC TCA CTA TAG GNNNNNNNNNNNNNNNNNNG TTT AAG AGC TAT GCT GGA AAC AGC ATA GCA AGT TTA AAT AAG GCT AGT CCG TTA TCA ACT TGA AAA AGT GGC ACC GAG TCG GTG CTT TTT TT-3′ containing a T7 promoter

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

(TAATACGACTCACTATAG), ~20-nt GCTTAGTTACGCGTGGACGA for transcription of Sc-gRNA or 18 bp sequences of VKG1-gRNA TCGCTGCGCTCGGTCGTT followed by gRNA scaffold. The DNA templates were generated individually *in vitro* using Gibson assembly (New England Biolabs) (Supplementary Table 12). The resulting dsDNA was then purified using NucleoSpin Gel and PCR Clean-Up (Takara) and approximately 200 ng of DNA was used as template for a T7 *in vitro* transcription (IVT) reaction (MAXIscript T7 Transcription Kit from Invitrogen) generating the gRNAs. In the next step, *in vitro* transcribed gRNAs were treated with DNase (DNaseI-Thermo) and purified using the Omega EZNA PF kit. The concentrations of the purified gRNAs were then measured using a NanoDrop. After purification, each gRNA was diluted to 200 ng/µL and stored at –80°C. Cleavage assays were conducted in a reaction volume of 20 µL containing SpCas9 (homemade) (200 nM), each gRNA (400 ng), each donor plasmid (1000 ng) in 1X cleavage assay buffer (homemade) at 37°C tested over the time periods of 2, 4 and 8 hours. The cleaved dsDNA was analyzed by 1.5% agarose gel electrophoresis (Fig. 2d).

Cell culture

HEK 293T cells were cultured in Dulbecco's modified Eagle's Medium (DMEM, Wisent) supplemented with 10% fetal bovine serum vol/vol) (FBS, Wisent) at 37 °C and 5 % CO_2.

Dose-response curve for antibiotic selection of mammalian cells

We measured the kill curve for each of the antibiotics to determine the minimum amount of the antibiotic required to eliminate non-infected or non-edited cells over the course of 7 days. This was accomplished by performing the assay using cell counting kit-8 (SIGMA). HEK 293T cells were seeded in 24-well plates 24 hours prior to antibiotic treatment. Cells were treated with Zeocin (Gibco) at concentrations ranging between 0 (negative control) and 500 µg/ml for 7 days. For Blasticidin S HCl (Thermo), cells were treated with concentrations between 0 (negative control) and 15 µg/ml for 7 days. At the end of the assay, cells were treated with Cell Counting Kit-8 (CCK-8, Sigma) according to manufacturer's instruction and an estimate of the number of surviving cells was made based on OD450 measurements (Extended Data fig. 3a, b).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Design and optimization of donor DNA knock-in based on NHEJ DNA repair

We aimed to develop an efficient method for inserting exogenous DNA fragments into the human genome by harnessing the endogenous non-homologous end joining (NHEJ) DNA repair pathway. NHEJ is the most efficient repair system in mammalian cells since it is functional during all phases of the cell cycle[8]. Our design consists of two main steps: First, simultaneous introduction of double-strand breaks (DSB) in both genomic and donor vectors inside living cells and second, integration of the linearized donor vector into the cleaved genome by NHEJ DNA repair system (Fig. 1a, b). To efficiently apply the NHEJ method, we optimized three individual plasmid elements: 1, our modified version of the SpCas9-NG plasmid (derived from pX330) that harbors Sc-gRNA or VKG1-gRNA, 2, The mini-circular donor vectors harboring coding sequences for the Zeocin resistance gene (Zeo) fused to a P2A self-cleavage peptide and flanked with complementary sequences of Sc-gRNA or VKG1-gRNA followed by either AGG or TGG as PAMs for the Sc-gRNA and CGG or TGG for VKG1-gRNA and 3, our modified version of lentiCRISPR-V2 -vector harboring locus specific gRNAs. To evaluate the efficiency of our NHEJ method for integration of donor DNA into the genome, we first targeted five human-specific genes including RPS6KB1, AKT1, PKD1, RPS6 and PPP2CA. These genes were chosen because first, they have diverse expression levels, which could be an ideal representative of the broad distribution of transcript levels for the human genome and second. Additionally, we have previously tagged these genes without functional consequences[22]. We made a lentiCRISPR library consisting of twenty-two locus specific gRNAs targeting above-mentioned genes using protocol established by Ran *et al.*[36] (SupplementaryTable1).

To test for integration efficiency, well-dissociated HEK 293T cells were seeded on 6-well plates (Corning) 24 h before transfection, at a density of 2×10^6 cells per well in a total volume of 2 ml of complete DMEM medium (DMEM, 10% FBS and 1X Sodium Pyruvate). Cells were transfected with a mixture of 3 µg plasmid coding SpCas9-NG and 3 µg of each donor-plasmids pre-mixed with X-tremeGene 9 transfection reagent (36 µl) (Roche) according to the manufacturer's protocol and infected at the same time with 200 µl of Lentiviral lentiCRISPR-22-gRNAs. As successful gene editing requires a minimum threshold expression level of Cas9, we also tested the integration efficiency of our model in cells that were transfected with SpCAS9-NG

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

plasmid 18 hours before introducing donor vectors and lentiCRISPR-22-gRNAs giving cells time to produce SpCAS9-NG. To select for transduced cells, in the next step, cells were treated with 4 µg/ml of Blasticidin S over 7 days. Five days after transduction, cells were selected following Zeocin treatment (100 µg/mL) until all the non-transduced cells and cells without successful integration were dead. At the end of the screening, cell viability was measured using Cell Counting Kit-8 (CCK-8, Sigma) and used to evaluate the percentage knock-in efficiency. Results were obtained from three replicate wells and presented as average efficiency ± standard error.

Designing two custom gRNA libraries for N- and C-termini tagging protein-coding sequences

For the selection of coding regions using the UCSC Table Browser[41], we downloaded the table "RefSeq Curated" from human assembly hg38 (accessed on 27/03/20), group "Genes and Gene Predictions" and track "NCBI RefSeq". We excluded non-coding entries from the table by filtering out rows with "NR_" accession prefixes. We also downloaded the "ccdsInfo" table (hg38 –> Genes and Gene Predictions -> CCDS). We created a new table by merging both tables on RefSeq accessions, thus linking protein-coding transcripts (NM) to their CCDS entry. Finally, we removed entries belonging to chr_*alt and chr_*fix. The final table contains 18,804 unique gene IDs and 38,788 unique RefSeq accessions (SupplementaryTable13). For the design of guide RNA libraries, we used the command-line version of CHOPCHOPv3[41] to find guides for our targets. We forked the source code from Github and implemented the following in-house features: Limiting the guide search to either of the terminal exons, detecting the frame of candidate guides and computing the number of residues removed from the original reading frame following non-homologous knock-in. These additions, combined with the removal of several features from the original program, allowed us to minimize the search space by pre-emptively excluding guides that failed to meet the criteria for successful knock-in. The search was limited to guides of length 20 nucleotides with the protospacer-adjacent motifs (PAMs) NGN, NAH and NTG. For each accession in our master table, we used the program to find guides that cut in-frame and trimmed at most three residues from either terminus. We further constrained guides to have a minimum GC content of 20% and a maximum of 85%. Guides with oligo(dT) tracts of four or more were excluded to avoid premature transcription termination by RNA Pol III[42.] Guides with restriction motifs "ACCGGT",

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

"GGTACC", "GAATGC" and "CGTCTC" were also excluded. These steps resulted in lists of 89,641 and 85,942 unique guides for 5' and 3' (N- and C-terminal protein fusions), respectively. Then, for the exclusion of guides based on target features, we used the "Retrieve/ID Mapping" tool of the UniProt43 database (accessed on 20/05/20) with our list of RefSeq accessions to obtain the corresponding UniProt entries. We kept the reviewed entries only and queried them for "Molecule processing" features using the Proteins API[44]. Specifically, we recorded signal peptide, transit peptide and pro-peptide annotations found at either terminal in a table (Supplementary 3Table 1). Using this table, we excluded all guides targeting terminal regions corresponding to either of the sequence features above. Finally, we removed duplicated guides to obtain 73,909 and 85,394 unique guides for N- and C-termini, respectively (Supplementary Table 2 and 3). Thirty-four genes had no guides passing these filters (Supplementary Table 15).

Genome-scale lentiviral gRNA construction

All 79-mer oligonucleotides on corresponding microarray chips for N- and C-terminal libraries (Custom Array) were amplified by PCR as a pool (Supplementary Table 11). The PCR products were then purified using NucleoSpin Gel and PCR Clean-Up (Takara) and cloned into a modified version of the lentiviral vector LentiCRISPR-V2 (Addgene Plasmid #52961), which contains the mCherry gene. The LentiCRISPR-V2 vector was digested with BsmBI-V2 (New England Biolabs) for 1 hour at 55°C and gel purified using the NucleoSpin Gel and PCR Clean-Up (Takara). In the next step, the purified library was sub-cloned into the digested lentiCRIPSRv2 vector at a ratio of 1:10 vector-to-insert molar ratio, using a Gibson assembly kit (New England Biolabs). The ligation reactions were precipitated using isopropanol and transformed into Endura Electrocompetent cells (Lucigen). To yield a 2000-fold representation of each library, 40 identical ligation reactions were pooled and purified, followed by 48 parallel transformations. Outgrowth media from transformations were pooled and plated onto fifty 500 cm2 LB-carbenicillin (100 mg/ml) agar plates and were incubated at 37°C for 12 hours. Colonies were scraped off the plates, pooled, and the plasmid DNA was extracted using the ZymoPURE II Plasmid Maxiprep Kit (Zymoreserach). The extracted plasmids for each library were subjected to PCR reactions using NEBNext high-fidelity polymerase (New England Biolabs) (Supplementary Table 15) and cleaned up using the

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Zymo-Spin V with Reservoir. NGS was performed on purified samples on Nextseq 150 cycles, Mid Output, v2.5 (Illumina). We aimed for at least 100 reads per unique gRNA.

Counting gRNAs from NGS data

To investigate the counts and distribution of gRNAs in the library using NGS data, we created a custom snakemake-based tool in python that processes NGS data, maps the reads to the gRNA library, and estimates evenness of the read counts[45]. In brief, reads were extracted from fastq files, mapped to the gRNA library of choice and regions of the reads mapped to the library were extracted and counted. We implemented two different approaches for this objective. In the first approach, NGS reads perfectly matching to gRNA library subjects were extracted using Mageck not allowing any mismatches[46]. In case of mismatches, sequences were scored as unmapped. In the second approach, user-defined mismatch thresholds were considered for mapping reads to the gRNA library to account for mutations in NGS data that may occur during library preparation, PCR amplification, or NGS read errors using CD-HIT[47]. This approach, however, requires all pairwise gRNAs to have maximum similarity less than the defined threshold because the robust mapping of the reads with mismatches to correct gRNA would not be feasible for highly similar gRNAs. Using either approach, the gRNA counts per gRNA were then analyzed and the evenness of distribution of gRNA counts was measured by Gini index implemented in Mageck.

Lentivirus production

gRNA library lentivirus is produced by co-transfection of lentiviral vectors pCMV-Delta 8.9 plasmid (packaging vector), pCMV-VSV-G (envelope vector) and lentiCRISPR-V2 library plasmid, using X-tremeGene 9 transfection reagent (Roche). Briefly, the well-dissociated HEK 293Tcells were seeded in T225 flasks 20–24 h before transfection at a density of 2 × 107 cells per flask in a total volume of 45 ml complete medium. Overall, four flasks were used for each library. On the day of transfection, cells with 80% confluency were transfected with a mixture of pCMV-

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Delta 8.9 plasmid (9 µg), pCMV-VSV-G (7.7 µg), lentiCRISPR-V2 plasmid library (12 µg), and X-tremeGene 9 (270 µl), according to the manufacturer's protocol. At 18 hours after transfection, the medium was replaced with pre-warmed complete DMEM medium. The expression level of mCherry was used as the control for estimating the transfection efficiency. The first Virus-containing medium was harvested two days after transfection, cellular debris was filtered out using Millipore's 0.45 µm Stericup filter unit and stored at -80°C. To obtain sufficient library coverage, virus particles were collected from four 225 cm2 flasks for each library.

Pooled genome-wide CRISPR knock-in in HEK 293T for integration of exogenous donor DNA fragments

To expand PRISM to the integration of P2A-Zeocin and Zeocin-P2A fragments to the entire human protein-coding genes, we harnessed NHEJ DNA repair on a large scale as follows:
A total of four 225-cm2 flasks each containing 2×10^7 HEK 293T cells were seeded 20 hours prior to transfection for 5' and 3' integration. On the day of transfection, each flask was transfected with donor cleaving vector (SpCas9-NG-Sc-gRNA). At 18 hours after transfection, cells were simultaneously infected with lentiCRISPR-gRNA library and mini-circular plasmid harboring P2A-Zeocin, Zeocin-P2A, or liker-mClover3-P2A-Zeocin donor vectors. The infection was carried out at a multiplicity of infection (MOI) of 0.3 to ensure that most cells receive only one gRNA copy. At 48 hours after infection, Cells were transferred to 500 cm2 plates and treated with Blasticidin S (4 µg/ml) for 7 days. Five days after infection/transfection, we started treating cells with Zeocin at 100 µg/mL over the course of two weeks to remove cells in which either no donor was inserted, a donor was inserted out of frame, or a product fusion protein was toxic to cells. Surviving cells were then pooled and stored at 500-fold coverage using Cell banker (AMSBIO). One sample of the cells with successful integration was subjected to fluorescent imaging and for each end (5'- or 3'-) integration, samples were subjected to genomic DNA extraction, PCR amplification and Illumina next-generation sequencing using Nextseq 150 cycles, Mid Output, v2.5 kit and aimed for 130 million reads in total (primers are listed in Supplementary Table 16).

mClover3 Fluorescence Imaging of the pooled library

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Forty-eight h prior to imaging, cells were grown at a density of 7 x 10^5 cells per dish in 35 mm Nunc™ Glass Bottom Dishes (ThermoFisher) in DMEM without phenol red. Cells were imaged on an inverted Nikon Ti2-equipped with 20x/0.5 Air Ph1 WD 2.1, 60x/1.4 Oil DIC WD 0.13, 100x/1.45 Oil Ph3 WD 0.13, 100x/1.45 Oil DIC WD 0.13, 4x/0.2 Air WD 20 , 20x/0.75 Air DIC WD 1.0.and 405 nm, 488 nm, 561 nm and 642 nm lasers. Images were taken using 488 nm laser at 50% power, 200ms exposure time, the 60x/1.4 Oil DIC objective, NIS Elements version 4.60 software and analyzed by Fiji software.

Genomic DNA extraction and genomic PCR

Genomic DNA was extracted from a sufficient number of cells to maintain coverage of >500 using the Zymo Research Quick-gDNA MidiPrep according to the manufacturer's protocol. In the next step, all genomic DNA harvested from the screen was PCR amplified using NEBNext high-fidelity polymerase (New England Biolabs) (Supplementary Table 16) and cleaned up using the Zymo-Spin V with Reservoir. NGS was performed on purified samples on Nextseq 150 cycles, Mid Output, v2.5 (Illumina); we aimed for 100 reads per individual gRNA.

mClover3 Fluorescence Imaging of single cells

Cells were incubated in a glass bottom 96-well plate (Optical CVG- ThermoFisher) and were imaged using a GE InCell Analyzer 6000.The imaging parameters were the following: excitation was performed using the 488nm 25mW laser diode at 100% with a 100ms time exposure. Separation between excitation and emission light was performed using a quad band-pass polychromic mirror BP 309/40 + BP 482/18 + BP 564/9 + BP 640/14. Emission light was filtered using a single band pass BP 525/20.One field per well was imaged using a 60x/0.95 Plan Apo air Objective from Nikon.Focus was done using a 785nm laser autofocus with 3um offset. Images were captured a 2048x2048 sCMOS camera with 16-bit depth and rolling shutter giving a final pixel size of 0.1083 µm^2.

Image processing of knock-in cells

Images were collected as TIF files. Images were segmented using manual intensity thresholding using ImageJ [48]. The mean intensity of the segmented area was measured on the original 16-bit

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

761 images. Because mClover3 intensity varies greatly between cell lines, brightness and contrast were
762 reset to display the min and max intensities for each image. 8-bit images were then generated and
763 inserted in the figure.

764 **Identification of donor DNA insertion site by nested PCR**

765

766 To identify the location of the donor DNA in the human genome, we used a modified version of a
767 method described by Lourdault, Matsunaga *et al.* to identify the insertion sites of transposon
768 mutants during golden Syrian Hamster infection[49]. Here, we adapted their protocol to identify the
769 location of our donor DNA fragments. For this objective, first the cells that have integrated donor
770 DNA fragments were subjected to Genomic (gDNA) extraction as described above, and sheared
771 by sonication into fragments of between 200 and 600 bp. Next, we used terminal deoxynucleotidyl
772 transferase (TdT) (New England Biolabs) to add repeats of CCCCCCCCCCCCCCCC (C-tail) at
773 the end of fragmented DNA in accordance with manufacturing protocol. The C-tailed genomic
774 library was subjected to two rounds of PCR (Fig. 4a and Supplementary Table 9). The first PCR
775 was performed with a forward primer which recognized a part of the donor DNA and the reverse
776 primer designed on the C-Tail (GGGGGGGGGGGGGGGG). Product of the first PCR was
777 subjected to the second round of PCR. This step enables the addition of the adapters necessary for
778 the sequencing. The resulting PCR products were subjected to PCR clean up (Takara) followed by
779 NGS using NextSeq 500 (Illumina) high output 2 x 75 bp Flow Cell.

780

781 **Processing Semi-nested NGS data**

782

783 Semi-nested PCR NGS data were pre-processed, trimmed, and mapped to the human genome in
784 multiple steps. First, forward reads that contained the plasmid sequence 'TCCACGCGTAA'
785 (Plasmid sequence search metrics: coverage: 100%, allowed mismatches: 1 bp) along with their
786 corresponding reverse reads were filtered out using cutadapt[50]. Next, the linker sequence
787 'GGTACCGGTGG' was trimmed from 5' ends of forward reads (linker sequence search metrics:
788 coverage: 100%, allowed mismatches: 1 bp) and at the same time the reads without the linker and
789 the corresponding reverse reads were filtered out. Additionally, the linker sequence
790 'GGGTCCCCATCG' was trimmed from 5' ends of forward reads (linker sequence search metrics:
791 coverage: 100%, allowed mismatches: 1 bp). In the next step, polyG tails from 5' ends of reverse

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

792 reads along Ns and low-quality bases (Q < 10) from both 5' and 3' ends of forward and reverse
793 reads were trimmed using cutadapt. The trimmed reads were then mapped to human genome
794 assembly hg38 (Downloaded from UCSC genome browser) using bowtie2[51] with options "--local"
795 and "--very-sensitive" and defining DNA fragment length of 50 bp as minimum and 700 bp as
796 maximum. The correct paired reads with forward and reverse reads mapped within the defined
797 DNA fragment length on the human genome were then sorted using samtools[52] and the read pairs
798 overlapping with genomic locations targeted by locus-specific-gRNAs in the library were
799 retrieved.
800

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Acknowledgments

The authors thank Professor James Omichinski and Truche Sébastien for providing cas9 protein used in *In vitro* Cas9 cleavage assay and Monique Vasseur and the Microscopy Platform form the faculty of medicine of Université de Montréal. The authors thank Caitlin Taylor of Illumina Inc. for her support in providing the next-generation sequencing kits used in this study. The authors acknowledge support from Canadian Institutes of Health Research (CIHR) grants MOP-GMX-152556 (SWM) and RNI00417 (AS) and training fellowships from FRQNT (LG) and Mitacs (MK).

Author Contributions

SWM and MK conceptualized this study. MK designed and optimized NHEJ-based knock-in module, constructed all the required plasmids for the study and acquired data for NHEJ efficiency, *In vitro* Cas9 cleavage assay, integration of exogenous donor DNA fragments into the genome and prepared samples for NGS of gRNA CRISPR libraries and semi-Nested PCR. LG and AS generated the pipeline for designing gRNA-CRISPR-libraries. MMS generated pipelines for analyzing NGS data for gRNA-CRISPR-libraries and semi-Nested PCR and analyzed the resulted data.SN processed data for protein localization and fluorescent intensity. SWM and MK wrote the manuscript and all authors provided feedback and editorial support.

Competing interest declaration

The authors declare no competing interest.

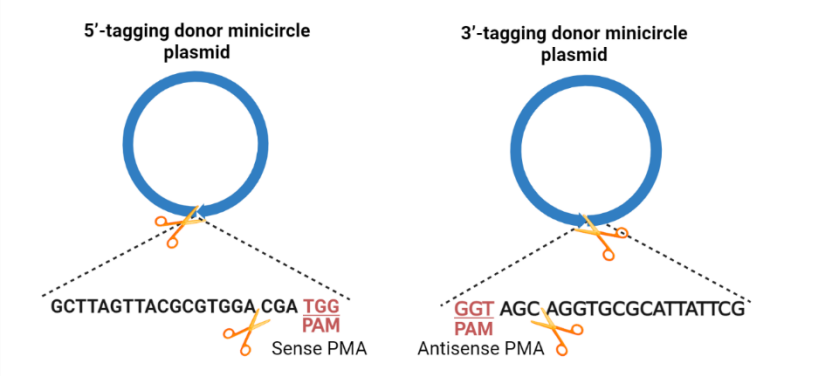
Data availability statement

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

All data generated or analyzed during this study are included in this manuscript (and its supplementary information files)

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Extended Data



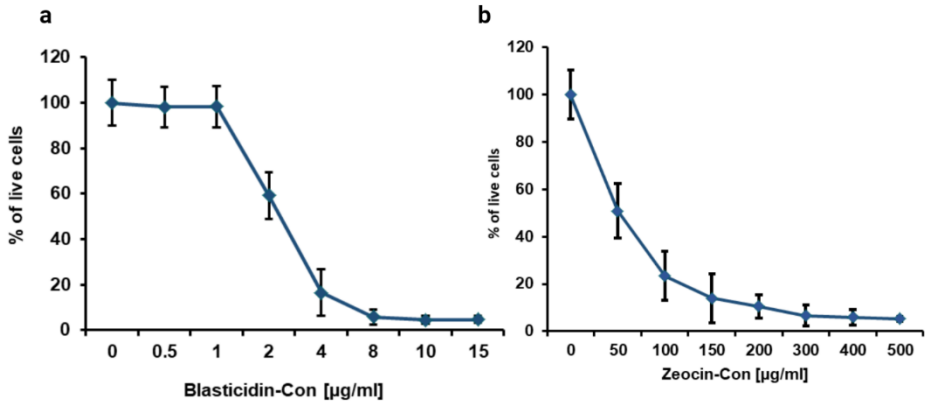
Extended Data Fig. 1 | MC-Sc-gRNA guide-directed SpCas9-NG PAM sites. The Sc-gRNA was chosen as guide for cleavage of the minicircle (MC) donor plasmids. There sequences were chosen because they do not exist in any protein-coding gene in the human genome. Two types of MC-Sc-gRNA PAM sites were created: donor DNA sequence for 5 prime tagging in which PAM is on the sense strand and donor DNA for 3' tagging in which the PAM is on the antisense strand.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



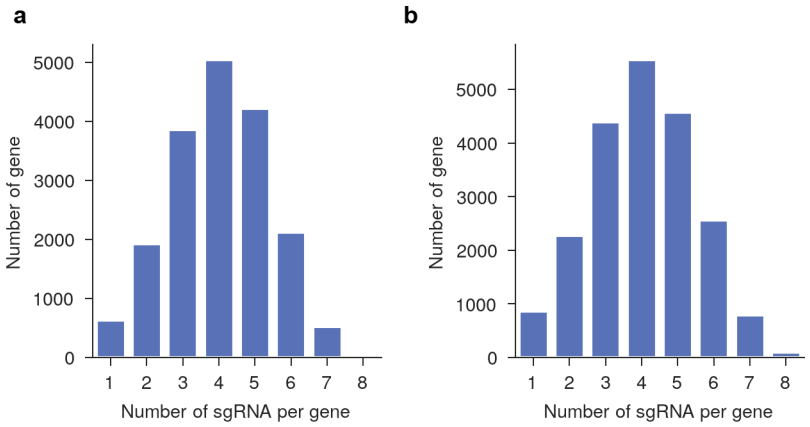
Extended Data Fig. 2 | The design of donor-cleaving vector and gRNA-CRISPR-Library. a, The scheme indicates the U6 promoter-based expression cassette followed by the Sc-gRNA (orange), gRNA scaffold (blue) and termination signal (red). **b**, The scheme indicates the U6 promoter-based expression cassette followed by the gRNAs targeting 18,804 human genes, including their splice variants (cyan), gRNA scaffold (blue), capture sequence (pink) and termination signal (red). The capture sequence is a 22 bp DNA fragment, which is designed to provide a priming site to later amplify the gRNA library from the Lenti-CRISPR plasmid and to help to distinguish the donor cleaving -gRNA from the gRNA library in later steps.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



Extended Data Fig. 3 | . Dose response curve for Blasticidin and Zeocin selection for HEK 293T cells. Cell viability assays were performed to determine the minimum amount of the Blasticidin and Zeocin required to eliminate non-infected / non-edited cells over 7 days. HEK 293T cells were seeded in 24-well plates 24 hours prior to antibiotic treatment. In the case of Blasticidin, cells were treated with concentrations of Blasticidin of between 0 (negative control) and 15 µg/ml and for Zeocin, cells were treated with concentrations of between 0 (negative control) and 500 µg/ml of Zeocin for 7 days. At the end of the screen, cells were treated with Cell Counting Kit-8 (CCK-8, Sigma) according to manufacturer's instruction and an estimate of the number of surviving cells was made based on OD_{450} measurements.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



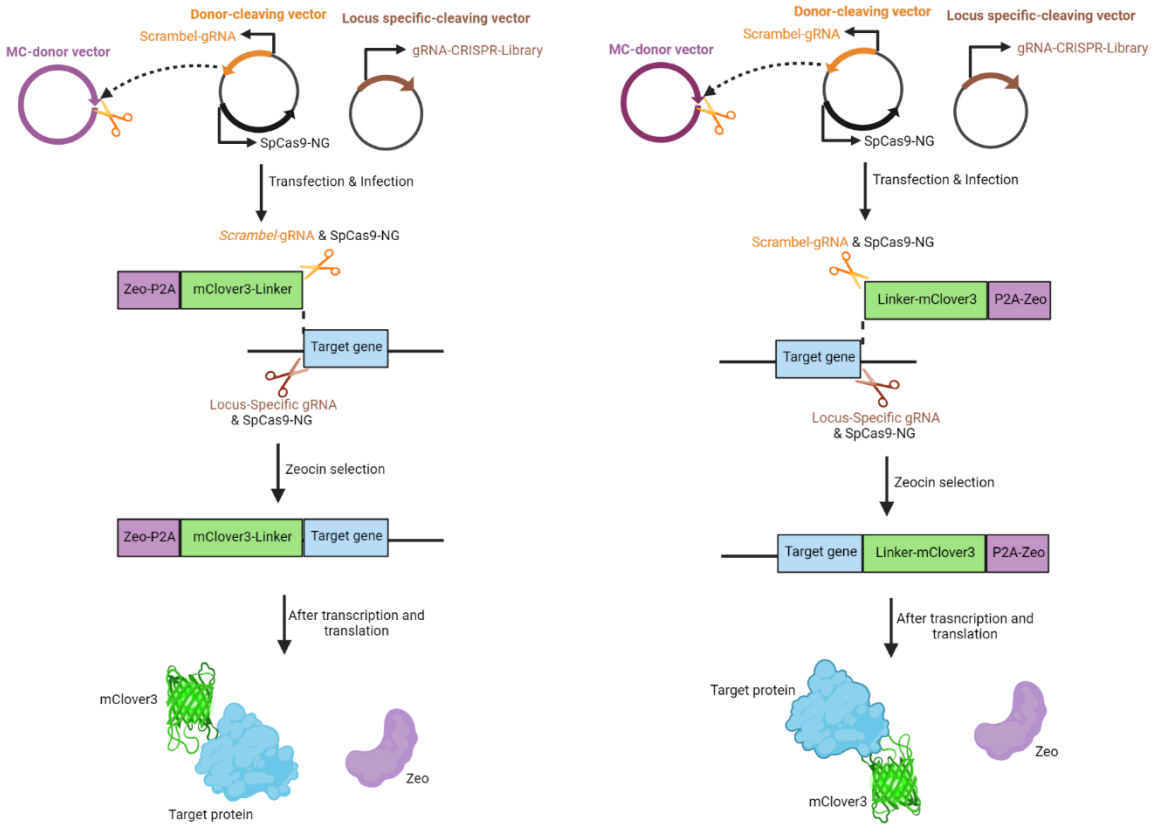
873

874 **Extended Data Fig. 4 | Two gRNA libraries consisting of minimum of 1 and maximum 8**
875 **gRNAs per splice variant of human genes were designed.** We designed a genome-wide resource
876 of two gRNA libraries consisting of gRNA that 1, target PAMs resulting in the only in-frame
877 double strand breaks; 2, minimal off-target Cas9 cleavage; 3, maximal on-target Cas9 cleavage, 4,
878 avoidance of homopolymer stretches (e.g., AAAA, GGGG) and 5, GC content between 20 to 80%.
879 Following the above criteria, our libraries consist of a minimum of 1 and maximum of 8 gRNAs
880 per splice variant of human genes.

881

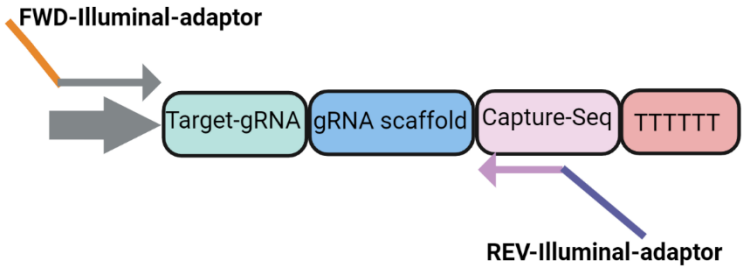
882

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



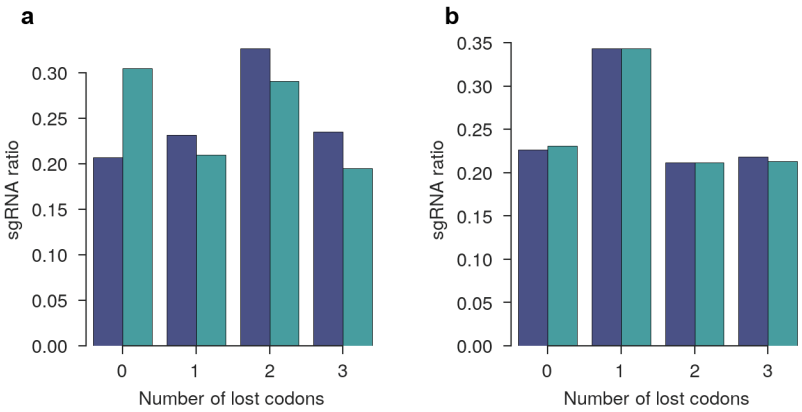
Extended Data Fig. 5 | Design and assessment of NHEJ-based knock-in modules for integrating mClover3 green fluorescent protein into the 5' and 3' end of target protein. Schematic of vectors and methods used to establish knock-in system for gene tagged with mClover3. The system comprises three vectors: a donor vector that contains coding sequences for peptide linker-mClover3 (green) followed by an antibiotic resistance gene as a selection marker fused to P2A self-cleaving peptide (purple), a donor cleavage vector that carries Sc-gRNA (orange) and SpCas9-NG (black), and a vector that cleaves a specific target gene in the genome (brown).

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



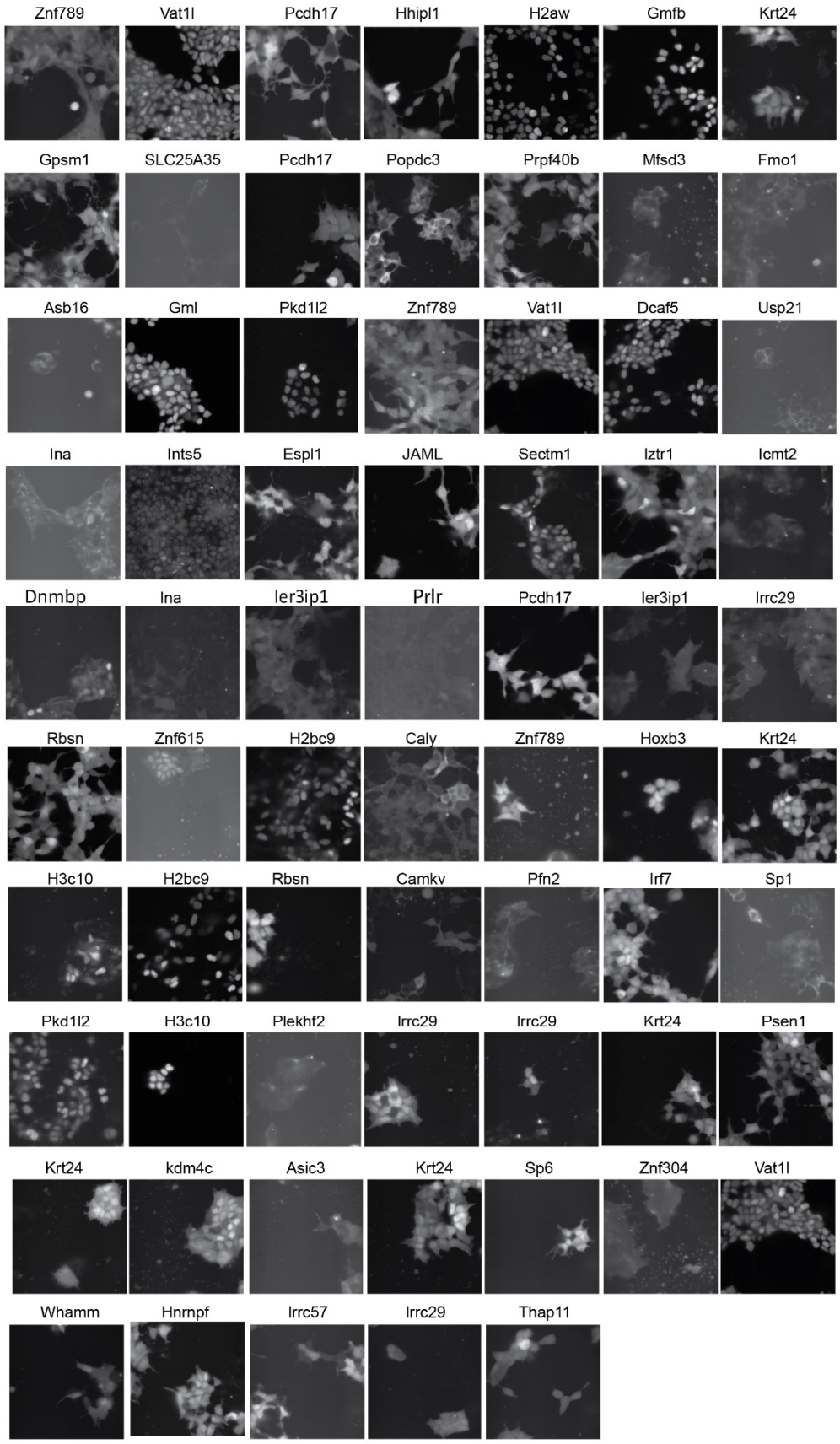
Extended Data Fig. 6 | The location of forward and reverse primers that were used to sequence the gRNA-CRISPR-library. Since both gRNA library and donor cleaving-gRNA have the same promoter and scaffold sequences, thus, we used a 20bp short sequence named ‘ capture sequence’ to provide a priming site for PCR amplification of gRNA-CRISPR library.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



Extended Data Fig. 7 | gRNA ratio corresponded to the gRNAs that resulted in the loss of 0 ,1 , 2 or 3 codons. As indicated here most of recovered gRNA sequences corresponded to the gRNAs that resulted in the loss of 0 or 1 codons, suggesting that integrations were selected for minimal loss of codons.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

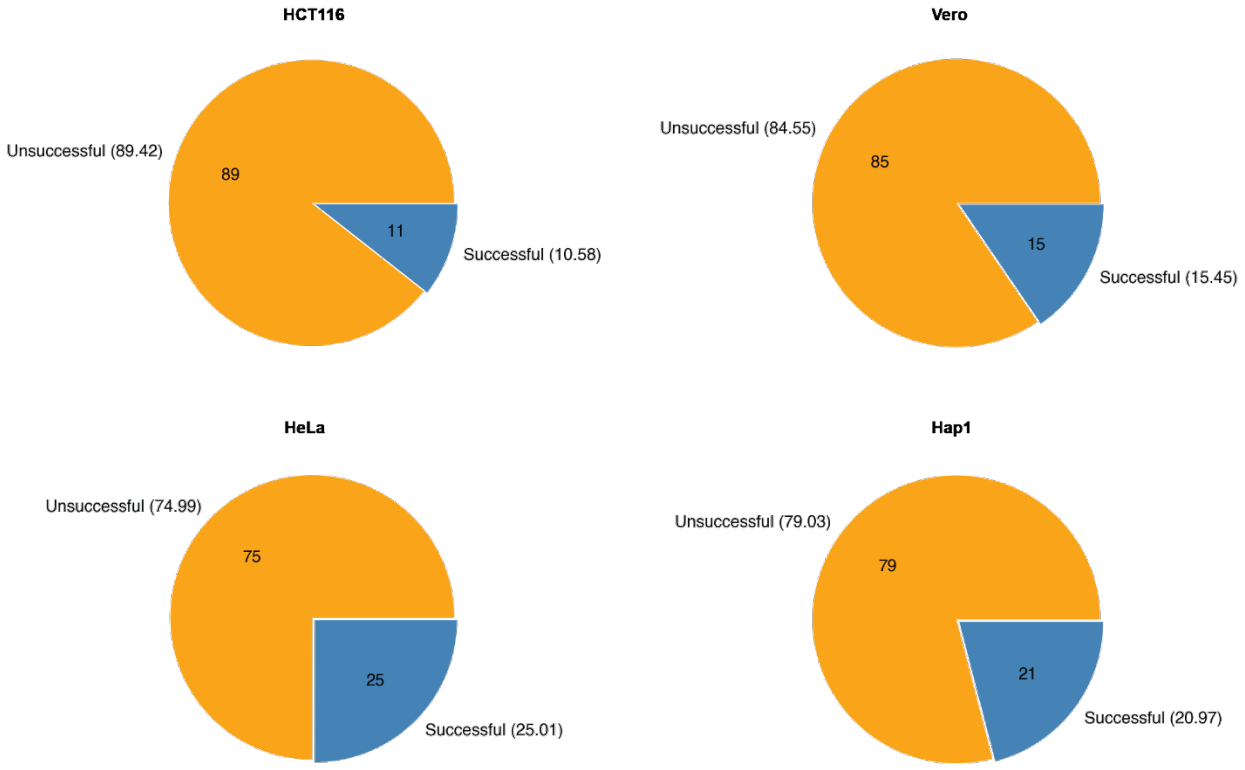


929

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

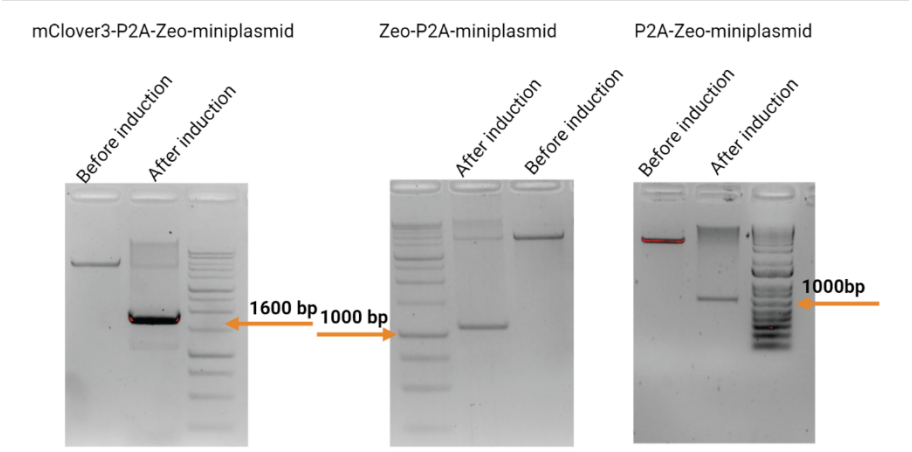
Extended Data Fig. 8 | Examples of successful targets showing fluorescent signal. *V*isualization of each single cell in which endogenous proteins are fused to mClover3 using our CRISPR/Cas9-NHEJ-based strategy. We observed variations in fluorescence intensity and patterns of expression, suggesting successful tagging of multiple proteins with varying abundances and different subcellular localizations. Since single cells were collected randomly, some proteins may appear more than once. However, if they were collected from the same plate, they were discarded from the analysis, assuming that they could come from the same original tagged cells. Conversely, if they were collected from different plates, they were kept in the analysis since they could be another confirmation of robust tagging, as they show the same localization and fluorescent intensity.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



Extended Data Fig. 9 | Genome-wide CRISPR integration of Reporter protein donor DNA fragments into Hap1, Hela, Vero-E6 and HCT116 cell lines. Our method demonstrates reasonable effectiveness in cell lines other than HEK293, although the efficiency varied based on the cell line.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.



969

970 **Extended Data Fig. 10 | Quality of mini plasmid extraction was confirmed by restriction**
971 **enzyme digestion.** 1.0 µg of minicircle plasmid before and after arabinose induction were digested
972 with SmaI restriction enzyme and then run on a 1.5% agarose gel to determine plasmid quality.
973 Parental plasmid indicates the size of plasmid before induction and minicircle shows the plasmid
974 size after induction. Successful minicircle plasmid extraction results in a band at 1600 bp.

975

976

977

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

978 **SUPPLEMENTARY TABLES**

979 **Supplementary Table 1** | gRNAs that were designed to target a set of five human-specific genes
980 including PS6KB1, AKT1, PKD1, RPS6 and PPP2CA at both 5' and 3' to protein-coding
981 sequences in HEK 293T cells.

oligoes	Assay	Sequences
5-sgRNA-RPS6KB1	Cloning of gRNA into LentiCRISPR-V2	CACCGCGCCTCCTTCGTCGCCTCAT
3-sgRNA-RPS6KB1	Cloning of gRNA into LentiCRISPR-V2	AAACATGAGGCGACGAAGGAGGCGC
5-sgRNA-RPS6KB1	Cloning of gRNA into LentiCRISPR-V2	CACCGGGCCCATGAGGCGACGAAGG
3-sgRNA-RPS6KB1	Cloning of gRNA into LentiCRISPR-V2	AAACCCTTCGTCGCCTCATGGGCCC
5-sgRNA-RPS6KB1	Cloning of gRNA into LentiCRISPR-V2	CACCGCCATGAGGCGACGAAGGAGG
3-sgRNA-RPS6KB1	Cloning of gRNA into LentiCRISPR-V2	AAACCCTCCTTCGTCGCCTCATGGC
5-sgRNA-RPS6KB1	Cloning of gRNA into LentiCRISPR-V2	CACCGCCTTCGTCGCCTCATGGGCC
3-sgRNA-RPS6KB1	Cloning of gRNA into LentiCRISPR-V2	AAACGGCCCATGAGGCGACGAAGGC
5-sgRNA-RPS6KB1	Cloning of gRNA into LentiCRISPR-V2	CACCGGCGGCGGGTCCGGGCCCATG
3-sgRNA-RPS6KB1	Cloning of gRNA into LentiCRISPR-V2	AAACCATGGGCCCGGACCCGCCGCC

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

5-sgRNA-RPS6KB1	Cloning of gRNA into LentiCRISPR-V2	CACCGCGCCTCCTTCGTCGCCTCAT
3-sgRNA-RPS6KB1	Cloning of gRNA into LentiCRISPR-V2	AAACATGAGGCGACGAAGGAGGCGC
5-sgRNA-AKT1	Cloning of gRNA into LentiCRISPR-V2	CACCGCTGGCCGAGTAGGAGAACTG
3-sgRNA-AKT1	Cloning of gRNA into LentiCRISPR-V2	AAACCAGTTCTCCTACTCGGCCAGC
5-sgRNA-AKT1	Cloning of gRNA into LentiCRISPR-V2	CACCGTAGGAGAACTGGGGGAAGTG
3-sgRNA-AKT1	Cloning of gRNA into LentiCRISPR-V2	AAACCACTTCCCCCAGTTCTCCTAC
5-sgRNA-AKT1	Cloning of gRNA into LentiCRISPR-V2	CACCGAGTGTGTGGACAGCGAGCGC
3-sgRNA-AKT1	Cloning of gRNA into LentiCRISPR-V2	AAACGCGCTCGCTGTCCACACACTC
5-sgRNA-AKT1	Cloning of gRNA into LentiCRISPR-V2	CACCGTGGCTATTGTGAAGGAGGGT
3-sgRNA-AKT1	Cloning of gRNA into LentiCRISPR-V2	AAACACCCTCCTTCACAATAGCCAC
5-sgRNA-PKD1	Cloning of gRNA into LentiCRISPR-V2	CACCGGCCCCAGCAGGACACCCCTT
3-sgRNA-PKD1	Cloning of gRNA into LentiCRISPR-V2	AAACAAGGGGTGTCCTGCTGGGGCC

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

5-sgRNA-PKD1	Cloning of gRNA into LentiCRISPR-V2	CACCGACCTTGTTCTTGGCCCGAAG
3-sgRNA-PKD1	Cloning of gRNA into LentiCRISPR-V2	AAACCTTCGGGCCAAGAACAAGGTC
5-sgRNA-PKD1	Cloning of gRNA into LentiCRISPR-V2	CACCGAAGGAGGACTAAGTGCTGCT
3-sgRNA-PKD1	Cloning of gRNA into LentiCRISPR-V2	AAACAGCAGCACTTAGTCCTCCTTC
5-sgRNA-RPS6	Cloning of gRNA into LentiCRISPR-V2	CACCGTTAGAAGTAGAAGCTCGCAG
3-sgRNA-RPS6	Cloning of gRNA into LentiCRISPR-V2	AAACCTGCGAGCTTCTACTTCTAAC
5-sgRNA-RPS6	Cloning of gRNA into LentiCRISPR-V2	CACCGCTGGATTCAGACTTAGAAGT
3-sgRNA-RPS6	Cloning of gRNA into LentiCRISPR-V2	AAACACTTCTAAGTCTGAATCCAGC
5-sgRNA-RPS6	Cloning of gRNA into LentiCRISPR-V2	CACCGCTACTTCTAAGTCTGAATCC
3-sgRNA-RPS6	Cloning of gRNA into LentiCRISPR-V2	AAACGGATTCAGACTTAGAAGTAGC
5-sgRNA-RPS6	Cloning of gRNA into LentiCRISPR-V2	CACCGGATTCAGACTTAGAAGTAGA
3-sgRNA-RPS6	Cloning of gRNA into LentiCRISPR-V2	AAACTCTACTTCTAAGTCTGAATCC

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

5-sgRNA-PPP2CA	Cloning of gRNA into LentiCRISPR-V2	CACCGGGCTCGCCTCTACGAGGTGC
3-sgRNA-PPP2CA	Cloning of gRNA into LentiCRISPR-V2	AAACGCACCTCGTAGAGGCGAGCCC
5-sgRNA-PPP2CA	Cloning of gRNA into LentiCRISPR-V2	CACCGACATGTGGCTCGCCTCTACG
3-sgRNA-PPP2CA	Cloning of gRNA into LentiCRISPR-V2	AAACCGTAGAGGCGAGCCACATGTC
5-sgRNA-PPP2CA	Cloning of gRNA into LentiCRISPR-V2	CACCGGGGGTACGACGAGTAACATG
3-sgRNA-PPP2CA	Cloning of gRNA into LentiCRISPR-V2	AAACCATGTTACTCGTCGTACCCCC
5-sgRNA-PPP2CA	Cloning of gRNA into LentiCRISPR-V2	CACCGTTTCATTACAGGAAGTAGTC
3-sgRNA-PPP2CA	Cloning of gRNA into LentiCRISPR-V2	AAACGACTACTTCCTGTAATGAAAC

Supplementary Table 2 (separate file) | List of gRNAs that were designed to target 18,804 human protein-coding gene sequences at their 5' end.

Supplementary Table 3 (separate file) | List of gRNAs that were designed to target 18,804 of human protein-coding gene sequences at their 3' end.

Supplementary Table 4 (separate file) | List of gRNAs that were recovered after cloning the gRNA library into LentiCRISPR-V2 to target 18,804 of human protein-coding gene sequences at their 5' end.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Supplementary Table 5 (separate file) | List of gRNAs that were recovered after cloning the gRNA library into LentiCRISPR-V2 target 18,804 of human protein-coding gene sequences at their 3' end.

Supplementary Table 6 (separate file) | List of gRNAs that were recovered after Zeocin treatment from cells with successful integration at the 5' end of human protein-coding gene sequences.

Supplementary Table 7 (separate file) | List of gRNAs that were recovered after Zeocin treatment from cells with successful integration at the 3' end of human protein-coding gene sequences.

Supplementary Table 8

Supplementary Table 9 | Primer sequences for direct verification of integration of donor DNA cassettes as fusions to protein-coding sequences and for semi-nested-PCR to confirm integration of donor DNA.

Primer name	Assay	Sequences
5' tagging-donor-DNA-FWD	semi-nested-PCR	GACTATCCCTGGCCTCCGCTCAGGAACGAGTCCAAGTACTTCCAAAGAATGACCACAACCTCTTCAGTGG
5' tagging-Ctail-REV	semi-nested-PCR	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGGGGGGGGGGGGGGG
5' tagging-Illumina-FWD	semi-nested-PCR	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC GATCTCAACCGGAATTGGG TACCGGTGGCGG
5' tagging-Illumina-REV	semi-nested-PCR	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

3' tagging-donor-DNA-FWD	semi-nested-PCR	AGAGGTTGTGGTCATTCTTTGGAAGTACTTGGA CTCGTTCCTGAGCGGAGGCCAGGGATAGTC
3' tagging-Ctail-REV	semi-nested-PCR	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATC TCCCCCCCCCCCCCCCC
3' tagging-Illumina-FWD	semi-nested-PCR	AATGATACGGCGACCACCGAGATCTACACTCTT TCCCTACACGACGCTCTTCC GATCTGTTCAATGGTCGAA CCATAGACCC
3' tagging-Illumina-REV	semi-nested-PCR	CAAGCAGAAGACGGCATACGAGATCGTGATGG TGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Supplementary Table 10 (separate file) | List of genes that were obtained from nested PCR and are tagged with donor DNA at the 5' or 3' end of human protein-coding gene sequences.

Supplementary Table 11 | Sequences for primers and DNA fragments that were used for preparing MC Donor vectors, donor-cleaving vectors, locus-specific- gRNA-CRISPR-library and preparing DNA templates for *in vitro* transcription of Sc-gRNA and VKG1-gRNA.

Primer name	Assay	Primer sequence
5' tagging -Zeocin-FWD	Cloning of Zeocin into pTubb3-MC	CCACAAGGAGACGACCTTCCAT GGCCAAGT TGACC
5' tagging-Zeocin-REV	Cloning of Zeocin into pTubb3-MC	AAGTTTGTTGCGCCGGATCCGTC CTGCTCCTCGGCC
3' tagging -Zeocin-FWD	Cloning of Zeocin into pTubb3-MC	TCGAAGAGAATCCTGGACCGAT GGCCAAGTTGACC
3' tagging -Zeocin-REV	Cloning of Zeocin into pTubb3-MC	CCAGTTGGGGCGGGCCCTCAGT CCTGCTCCTCGGCC

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

5'-P2A-Fragment	Cloning of P2A into pTubb3-MC plasmid	GCAACAAACTTCTCTCTGCTGAAACAAGCCGGAGATGTCGAAGAGAATCCTGGACCG
3'-P2A-Fragment	Cloning of P2A into pTubb3-MC plasmid	CGGTCCAGGATTCTCTTCGACATCTCCGGCTTGTTTCAGCAGAGAGAAGTTTGTTGC
mClover3- FWD	Cloning of mClover3 into pTubb3-MC	ATGGTGAGCAAGGGCGAGGAGCTGTTCACC
mClover3- REV	Cloning of mClover3 into pTubb3-MC	AAGTTTGTTGCGCCGGATCCCTTGTACAGCTCGTCCATGC
mCherry- FWD	Cloning of mCherry into LentiCRISPR-V2	GGTTCTAGAGCGCTGCCACCATGGTGAGCAAGGGCG
mCherry- REV	Cloning of mCherry into LentiCRISPR-V2	AAGTTTGTTGCGCCGGATCCCTTGTACAGCTCGTCC
CRISPR-Library-Forward	Amplifying Lenti-CRISPR-Library	GTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACC
CRISPR-Library-Reverse	Amplifying Lenti-CRISPR-Library	ACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC
5'-Sc-gRNA	Cloning of Sc-gRNA into PX330	CACCGGCTTAGTTACGCGTGGACGA
3'-Sc-gRNA	Cloning of Sc-gRNA into PX330	AAACTCGTCCACGCGTAACTAAGCC
5'-VKG1-gRNA	Cloning of VKG1-gRNA into PX330	CACCGCGCTGCGCTCGGTCGTT

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

3'-VKG1-gRNA	Cloning of VKG1-gRNA into PX330	AAACAACGACCGAGCGCAGCGAC

Supplementary Table 12 | Sequences of the synthesized oligoes that were used for cloning of Capture Seq after gRNA scaffold and preparing DNA templates for *in vitro* transcription of Sc-gRNA and VKG1-gRNA.

Synthesized oligoes	Assay	Sequences
5'-gRNA scaffold-Capture seq	Cloning of capture seq into LentiCRISPR-V2	GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCGCTTTAAGGCCGGTCCTAGCAA
3'-gRNA scaffold-Capture seq	Cloning of capture seq into LentiCRISPR-V2	TTGCTAGGACCGGCCTTAAAGCGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC
T7-promoter-Sc-gRNA-gRNA scaffold	*In vitro* transcription of SC-gRNA	TAA TACGACTCA CTATAG G GCTTAGTTACGCGTGGACGAG TTTAA GAGCAT GCTGGA AAC AGC ATA GCA AGT TTA AAT AAG GCTAG CCGTTA TCAACTTGA AAA AGT GGC ACC GAG TCG GTG CTT TTT TT
T7-promoter-VKG1-gRNA-gRNA scaffold	*In vitro* transcription of VKG1-gRNA	TAATACGACTCA CTATAGGTCGCTGCGCTCGGTCGTTGTTTAAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTA TCA ACT TGA AAA AGT GGC ACC GAG TCG GTG CTT TTT TT

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Supplementary Table 13 (separate file) | List of 38,788 unique accessions for the designed libraries.

Supplementary TableS14 (separate file) | list of signal peptide, transit peptide and pro-peptide annotations found at either N- or C- termini for the designed libraries.

Supplementary Table 15 (separate file) | List of 34 genes that had no guides in our designed libraries.

Supplementary Table 16 | Primer sequences for amplifying gRNA-CRISPR-library and NGS.

Primer name	Assay	Sequences
NGS-Fwd-1	gRNA library	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTAAGTAGAGGCTTTATATATCTTGTGGAAAGGACG
NGS-Fwd-2	gRNA library	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATCATGCTTAGCTTTATATATCTTGTGGAAAGGACG
NGS-Fwd-3	gRNA library	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGATGCACATCTGCTTTATATATCTTGTGGAAAGGACG
NGS-Fwd-4	gRNA library	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATTGCTCGACGCTTTATATATCTTGTGGAAAGGACG
NGS-Fwd-5	gRNA library	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCGATAGCAATTCGCTTTATATATCTTGTGGAAAGGACG
NGS-Fwd-6	gRNA library	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTATCGATAGTTGCTTGCTTTATATATCTTGTGGAAAGGACG
NGS-Fwd-7	gRNA library	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGATCGATCCAGTTAGGCTTTATATATC

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

		TTGTGGAAAGGACG
NGS-Fwd-8	gRNA library	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCGATCGATTTGAGCCTGCTTTATATATCTTGTGGAAAGGACG
NGS-REV-1	gRNA library	CAAGCAGAAGACGGCATACGAGATTCGCCTTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCGACTCGGTGCCACTTTTTTCAA
NGS-REV-2	gRNA library	CAAGCAGAAGACGGCATACGAGATATAGCGTCGTGACTGAGTTCAGACGTGTGCTCTTCCGATCTCCGACTCGGTGCCACTTTTTCAA
NGS-REV-3	gRNA library	CAAGCAGAAGACGGCATACGAGATGAAGAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCGACTCGGTGCCACTTTTTCAA
NGS-REV-4	gRNA library	CAAGCAGAAGACGGCATACGAGATATTCTAGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCGACTCGGTGCCACTTTTTCAA

1024

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

SUPPLEMENTARY SEQUENCES

U6-Sc-gRNA

>GAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGA
GAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTAGTACAAAATACGTG
ACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAAT
GGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATC
TTGTGGAAAGGACGAAACACCGGCTTAGTTACGCGTGGACGAGTTTTAGAGCTAGA
AATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGT
CGGTGCTTTTTTGTTTT

U6-CRISPR-gRNA-Library

>GAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGATACAAGGCTGTTAGA
GAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTAGTACAAAATACGTG
ACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAAATTATGTTTTAAAAT
GGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATTTCTTGGCTTTATATATC
TTGTGGAAAGGACGAAACACCGGNNNNNNNNNNNNNNNNNNGTTTTAGAGCTAG
AAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAG
CGGTGCGCTTTAAGGCCGGTCCTAGCAATTTTTTGTTTT

Zeo-P2A donor DNA

>ATGGCCAAGTTGACCAGTGCCGTTCCGGTGCTCACCGCGCGCGACGTCGCCGGAGC
GGTCGAGTTCTGGACCGACCGGCTCGGGTTCTCCCGGGACTTCGTGGAGGACGACTT
CGCCGGTGTGGTCCGGGACGACGTGACCCTGTTCATCAGCGCGGTCCAGGACCAGG
TGGTGCCGGACAACACCCTGGCCTGGGTGTGGGTGCGCGGCCTGGACGAGCTGTAC
GCCGAGTGGTCGGAGGTCGTGTCCACGAACTTCCGGGACGCCTCCGGGCCGGCCAT
GACCGAGATCGGCGAGCAGCCGTGGGGGCGGGAGTTCGCCCTGCGCGACCCGGCCG
GCAACTGCGTGCACTTCGTGGCCGAGGAGCAGGACGGATCCGGCGCAACAAACTTC
TCTCTGCTGAAACAAGCCGGAGATGTCGAAGAGAATCCTGGACCG

P2A-Zeo donor DNA

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.04.641506; this version posted March 11, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

>GGATCCGGCGCAACAAACTTCTCTCTGCTGAAACAAGCCGGAGATGTCGAAGAGA
ATCCTGGACCGATGGCCAAGTTGACCAGTGCCGTTCCGGTGCTCACCGCGCGCGACG
TCGCCGGAGCGGTCGAGTTCTGGACCGACCGGCTCGGGTTCTCCCGGGACTTCGTGG
AGGACGACTTCGCCGGTGTGGTCCGGGACGACGTGACCCTGTTCATCAGCGCGGTCC
AGGACCAGGTGGTGCCGGACAACACCCTGGCCTGGGTGTGGGTGCGCGGCCTGGAC
GAGCTGTACGCCGAGTGGTCGGAGGTCGTGTCCACGAACTTCCGGGACGCCTCCGG
GCCGGCCATGACCGAGATCGGCGAGCAGCCGTGGGGGCGGGAGTTCGCCCTGCGCG
ACCCGGCCGGCAACTGCGTGCACTTCGTGGCCGAGGAGCAGGACTGAG

Linker-mClover3-P2A-Zeo

GGCGGTGGCGGATCAGGAGGCGGTGGGTCTATGGTGAGCAAGGGCGAGGAGCTGTT
CACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGT
TCAGCGTCCGCGGCGAGGGCGAGGGCGATGCCACCAACGGCAAGCTGACCCTGAAG
TTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCTTC
GGCTACGGCGTGGCCTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTC
TTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTCTTTCAAGGAC
GACGGTACCTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAA
CCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACA
AGCTGGAGTACAACTTCAACAGCCACAACGTCTATATCACGGCCGACAAGCAGAAG
AACGGCATCAAGGCTAACTTCAAGATCCGCCACAACGTTGAGGACGGCAGCGTGCA
GCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGC
CCGACAACCACTACCTGAGCCATCAGTCCAAGCTGAGCAAAGACCCCAACGAGAAG
CGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATTACACATGGCATG
GACGAGCTGTACAAGGGATCCGGCGCAACAAACTTCTCTCTGCTGAAACAAGCCGG
AGATGTCGAAGAGAATCCTGGACCGATGGCCAAGTTGACCAGTGCCGTTCCGGTGC
TCACCGCGCGCGACGTCGCCGGAGCGGTCGAGTTCTGGACCGACCGGCTCGGGTTCT
CCCGGGACTTCGTGGAGGACGACTTCGCCGGTGTGGTCCGGGACGACGTGACCCTG
TTCATCAGCGCGGTCCAGGACCAGGTGGTGCCGGACAACACCCTGGCCTGGGTGTG
GGTGCGCGGCCTGGACGAGCTGTACGCCGAGTGGTCGGAGGTCGTGTCCACGAACT
TCCGGGACGCCTCCGGGCCGGCCATGACCGAGATCGGCGAGCAGCCGTGGGGGCGG
GAGTTCGCCCTGCGCGACCCGGCCGGCAACTGCGTGCACTTCGTGGCCGAGGAGCA
GGACTGA.