The Knock-In Atlas: A web resource for targeted protein trap by CRISPR/Cas9 in human and mouse cell lines

Yuma Hanai[1,4], Patrick Louis Lagman Hilario[1], Yuriko Shiraishi[1], Norisuke Kano[2], Minami Kojima[2], Kokoro Murai[2], Taro Kawai[2,3], and Katsutomo Okamura[1,4]

[1]Laboratory of RNA Molecular Medicine, Nara Institute of Science and Technology, 8916-5 Takayama Ikoma Nara 630-0192, Japan
[2]Laboratory of Molecular Immunobiology, Nara Institute of Science and Technology, 8916-5 Takayama Ikoma Nara 630-0192, Japan
[3]Life Science Collaboration Center (LiSCo), Nara Institute of Science and Technology, 8916-5 Takayama Ikoma Nara 630-0192, Japan

[4]Authors for correspondence
email: okamurak@bs.naist.jp or hanai.yuma.hv0@bs.naist.jp
TEL: +81-743-72-5410

Abstract

Various cell engineering techniques have been developed by leveraging the CRISPR-Cas9 technology, but large-scale resources for targeted gene knock-in are still limited. Here we introduce a tool kit for tagging genes by inserting artificial exons encoding fluorescent protein tags in target gene introns. To produce knock-in cells efficiently and reproducibly, we carefully chose and catalogued guide RNAs (gRNAs) for targeting genes in the human and mouse genomes by taking the gRNA efficacy scores and protein structures around the insertion sites into account. So far, we have constructed 427 gRNA expression plasmids to target 178 genes as the first set. The transfection and flow cytometry protocols were optimized for several cell lines including HEK293T, eHAP1, HeLa, THP-1, Neuro2a, mouse embryonic fibroblast (MEF) and mouse embryonic stem cell (mESC). A website has been launched to organize the results of initial characterization including flow cytometry data after transfection, confocal microscopy, and western blot results for the genes for which knock-in HEK293T cell lines were already made. We provide a user-friendly database to organize the information of the cell line and pre-designed gRNAs at <https://yumahanaiatokamuralab.shinyapps.io/KnockInAtlas/>.

Introduction

Gene knock-in technologies have evolved to satisfy various demands in molecular biology, from manipulation of gene sequences to tagging target genes with epitope, fluorescent protein or enzyme tags by using various of CRISPR/Cas9 technologies (Banan, 2020). Among the CRISPR/Cas9-based knock-in techniques, those utilizing non-homologous end-joining (NEHJ) hold the potential for large-scale cell line production because a single donor sequence could be used for insertions at various sites unlike the technologies utilizing the homology-directed repair (HDR) (Manna et al., 2019; Sawatsubashi et al., 2018; Schmid-Burgk et al., 2016). The issues that arise from the frequent introduction of mutations at junction by NHEJ could be circumvented by inserting a synthetic donor, where an artificial exon encoding a tag protein sequence is flanked by artificial splice acceptor and donor sequences, in the target gene intron to produce in-frame fusion proteins (Reicher et al., 2024; Serebrenik et al., 2019; Zhong et al., 2021). Because the mutations would not affect splicing patterns unless deletions reach the splice donor or acceptor sequences or insertions accidently create a new splice donor or acceptor. Indeed, this principle was used to produce a large scale knock-in cell library in a recent paper (Reicher et al., 2024).

Despite the available techniques and resources, there remain challenges in creating large sets of knock-in cell lines such as: 1) Setting up knock-in protocols optimized for different cell lines, 2) Designing effective gRNAs, 3) Selecting tagging sites to avoid disrupting protein structures, 4) Selecting genes that are expressed at reasonable levels.

Here, we provide a tool kit for NHEJ knock-in wherein a large set of gRNAs have been designed against human and mouse introns on a genome-wide scale while avoiding introns flanked by exons encoding protein domains with predictable rigid structures. A database of pre-designed gRNAs that is searchable by target gene symbols has been developed and each entry is associated with the TPM values of the host gene in common cell lines and the AlphaFold2 pLDDT scores of the 3 amino acids flanking the predicted tag-insertion site. As an initial set, we release 116 constructs whose successful knock-in was verified in HEK293T cells using the Venus knock-in donor and have made the gRNA information and initial characterization data of knock-in cells publicly available.

Results and Discussion

Designing targeting donors

In order to produce a knock-in library, we first considered several different forms of donor DNA. For simple detection of knock-in events, we used the Venus fluorescent tag that is flanked by splice acceptor and donor sites (Figure 1A left). Introns belong to one of three phases, intron (0), intron (1) and intron (2), where introns are located before 1st nucleotide, between 1st and 2nd and between 2nd and 3rd nucleotides in the codon, respectively. In the human and mouse genomes, the three intron phases, intron (0), intron (1) and intron (2) are present at a ~5:3:2 ratio in the CDS, respectively (Figure S1). To account for the reading frames in different intron phases, we prepared donor DNA with 0, 1 or 2 nt insertions (Figure 1A, Supplementary Table 1). Such constructs have been previously successfully utilized for NHEJ-mediated gene tagging by using donor plasmid DNA that was designed to be cleaved by a gRNA to release the exon cassette in cells (Reicher et al., 2020; Serebrenik et al., 2019; Zhong et al., 2021). To test other forms of donor DNA, we used PCR products with various 5' modifications to minimize immunostimulation and exonuclease-mediated degradation, and compared them with the conventional plasmid-based method (Figure 1B).

We tested the four forms of donor DNA to insert them in intron1 of the HNRNPA1 locus in HEK293T cells (Figure 1C). As a positive control, we transfected cells with the plasmid as a donor along with a plasmid expressing a gRNA to release donor DNA, in addition to a plasmid for the gRNA against HNRNPA1-Intron1 similar to the previously established strategy (Serebrenik et al., 2019). This method achieved the Venus positivity rate of ~1.4%, indicating successful knock-in at detectable frequency (Figure 1D and S2A). When cells were transfected with the donor PCR product without 5' modifications, we observed about 5% of "Venus positive" cells even when the Cas9-gRNA plasmid was omitted, and we assumed these cells to be damaged cells which often produce autofluorescence. Chemical modifications either in the 5' region at the phosphodiester bridges (phosphorothioate bridges at first 2 nucleotides: 2PS) or the 5' modifications linked with a C6 chain resulted in partial (2PS) or almost complete (5' modifications) suppression of the appearance of presumable autofluorescent cells (Figure 1D, Figure S2A), similar to the observations seen with HDR-mediated knock-in experiments (Yu et al., 2020). Various Venus positivity rates (3-4%) were observed with the PCR products with the chemically

modified primers at the first sorting 1 weeks after transfection, and became > 90% Venus positive populations after the amplification of sorted cell populations, except the sample transfected with the unmodified PCR product (Figure 1E, Figure S2B). Knock-in using the unmodified PCR products failed to produce a population with a high percentage of Venus positive cells after expanding sorted cells. This might be attributable to the high percentage of presumable autofluorescent cells that interfered with the enrichment of Venus positive cells at the initial cell sorting. While the mechanism behind this phenomenon remains unclear, our results indicated that 5' modifications, especially those with C6 or other additional modifications could be used as donor molecules, which may be especially beneficial when working with cell lines whose transfection efficiency is highly sensitive to the amount of DNA. The use of PCR products allows us not only to eliminate the unnecessary plasmid backbone of the donor, but also to omit gRNA against the donor plasmid.

We tested if this approach produces similar results in other cell lines. For this purpose, we used two human (eHAP1 and HeLa cells) and two mouse (Neuro2a and mESC) to perform similar experiments (Figure S3A). Interestingly, the relative Venus positivity rates with respect to the plasmid transfection were variable depending on the cell line, ranging from ~1.74-fold enhancement in HeLa cells to very low in mESCs. Nevertheless, Venus positive cells could be collected by sorting (Figure S4). Therefore, the effectiveness of the method with modified PCR products appeared to be variable depending on the cell line or perhaps transfection conditions. Therefore, we recommend the plasmid donor as the first choice, while C6-modified PCR products may be tested when plasmid donors produced poor results. Using the HNRNPA1-Intron1 gRNA, we tried our plasmid-based method on a human monocyte cell line THP-1 and mouse embryonic fibroblast (MEF) cells, which resulted in the successful production of Venus positive cells at the rate of ~1% and ~3%, respectively, further validating the application of the method in a broad range of cell types (Figure S3B).

We further tested whether this strategy could produce frequent random insertions occurring independently of the gRNA-mediated genomic DNA cleavage events. We employed splinkerette PCR to amplify the sequences surrounding the inserted Venus sequence (Potter and Luo, 2010). We chose 5 independent HNRNPA1-Intron1 knock-in eHAP1 lines obtained with the C6 modified PCR-based method (Figure S5). The splinkerette PCR products from 5 out of 5 clones corresponded to the HNRNPA1-Intron1 sequence, however, an

additional band was produced in one of the clones (Figure S5C), whose sequence indicated the insertion of Venus in the opposite direction, a potential drawback from our current NHEJ-based method. eHAP1 cells are known to spontaneously diploidize, and the knock-in event might have occurred in such diploid cells to produce two alleles with Venus insertions in both directions. Nevertheless, the results suggested that our NEHJ-based method does not cause frequent random integration of donor DNA. The actual random integration frequency remained unclear as only a small number of clones were tested, and the absence of randomly inserted copies needs to be tested when the applications require knock-in lines free of extra insertions of the tag. We also note that these results do not exclude the possibility of off-target insertions when gRNAs cause off-target cleavages.

Designing gRNAs

Various factors affect the knock-in rates, among which gRNA design is presumably most important. Good gRNAs to efficiently produce high-quality knock-in cells would satisfy the following criteria: 1) The gRNA mediates robust and specific cleavage of genomic DNA at the target site, 2) The target intron is efficiently spliced, 3) The insertion of the tag does not strongly interfere with the target protein function. In addition, 4) Lowly expressed genes would typically be excluded because the fusion protein needs to be endogenously expressed at detectable levels for subsequent experiments. The following subsections explain how our pipeline designs gRNAs satisfying each of the criteria.

1) Efficiency and specificity of gRNA cleavage

The cleavage efficiencies of gRNAs for all human and mouse introns of each representative transcript were predicted based on the machine-learning model developed by Doench et al., and deployed in the CHOPCHOP platform (Doench et al., 2016; Labun et al., 2019) (Figure 2A). gRNAs that have potential off-targets with no mismatch were omitted, and those with GC contents within 40-70% were selected. Up to three gRNAs for each intron with the highest cleavage scores were listed.

2) Selecting efficiently spliced introns

As mammalian genes produce multiple isoforms by alternative transcription start/termination and splicing, it is vitally important to select introns that are efficiently spliced for inserting our knock-in donor. A reasonable solution was to select a representative transcript isoform for each gene, which was catalogued

as canonical isoforms in The Matched Annotation from NCBI and EMBL-EBI (MANE v.1.0) (Morales et al., 2022) and in RefSeq Select for human and mouse, respectively (Goldfarb et al., 2025). In addition, we excluded the proximal and distal 75 nt of each intron from gRNA prediction, as they are within exon/intron junctions that often contain sequences important for splicing (Wachutka et al., 2019).

3) Avoiding rigid protein structures

To avoid interfering with the functions of the target protein, a plausible way is to avoid ordered regions. The pLDDT scores produced by AlphaFold2 (Jumper et al., 2021) are a promising genome-wide resource to predict such regions as tightly folded structures tend to be predicted more confidently (Binder et al., 2022). The predicted pLDDT scores were mapped to the human and mouse genomes at the codons encoding the corresponding protein sequence, and the average of pLDDT scores within the 3 amino acids surrounding the predicted tag insertion site was calculated (Figure 2B). The default pLDDT score threshold was set at 80 according to a previous study (Binder et al., 2022) but the users can modify the threshold using our interface (see below). With the cutoff of the pLDDT score at 80, there were 52,534 and 41,693 introns of 12,009 and 10,090 genes, in the human and mouse genomes, respectively (Figure 2C).

4) Expression level of the target gene

As the expression level of the tagged protein in knock-in cells is determined by the endogenous mechanisms via transcriptional and post-transcriptional regulatory elements, consideration of endogenous expression levels of its native counterpart protein is convenient for detecting fluorescently-tagged proteins after knock-in. For the practical application of knock-in cells, the tagged proteins need to be detectable by fluorescence from the tag (in the case of flow cytometry and cell imaging) or antibody-based detection (in the case of western blotting and immunostaining). As a reasonable approximation, we used mRNA-seq quantification to represent the gene expression level of each gene in the subject cell type. We used the transcriptome quantification in 4 human cell lines and 2 mouse cells that was made available by The Human Protein Atlas (Jin et al., 2023) and Expression Atlas (George et al., 2024; Lienert et al., 2011), respectively. In our interface, the TPM values for HEK293, HeLa, eHAP1, and THP-1 for human cells and, MEF and mESC for mouse cells are provided and users can set the threshold TPM values when selecting their gene of interest. In our experience with HEK293T cells, genes with the threshold of normalized TPM

(Set at > ~45 by default, but can be adjusted using the slide bar in the case of lowly expressed gene) would produce reporter lines with fluorescence readily detectable by flow cytometry (Figure 2D, Figure S7A) or fluorescent microscopy.

Web resource

To satisfy demands of individual researchers to predict good gRNAs for knock-in experiments, a user-friendly interface has been launched. This website is mainly consisted of three pages, “HEK293T knock-in cell lines”, “Human gRNA database” and “Mouse gRNA database”. In the “HEK293T knock-in cell lines” page, the main table shows a list of knock-in lines, and contains an experimental summary to indicate whether flow cytometry data after knock-in transfection, fluorescent microscopy and Western blotting results are available (Figure 3A). The table contains two URL links, one of which prompts a new page to show our experimental results and the other to the webpage of The Human Protein Atlas of the corresponding gene (Thul et al., 2017) so the users can easily compare the results for the same gene in the two resources. Our experimental summary page shows the results with HEK293T cells including flow cytometry data of Venus intensity 2 days after transfection, confocal microscopic images of bulk knock-in cells except for HNRNPA1 with Hoechst staining and Western blot gel pictures of knock-in lines derived from single cells using an anti-GFP antibody against the Venus tag (Figure 3B).

The second and third pages show the pre-designed gRNA information for the human and mouse genomes, respectively. There are three parts with two sortable tables and a JBrowse window (Figure 4). The first table is searchable by gene symbol or for representative gene functions (Figure 4A). When a row of the first table is selected, the user will see a list of introns within the CDS of the selected gene in the second table (Figure 4B) and the JBrowse window will show the selected gene locus (Figure 4C). The JBrowse window has a gene annotation track with a representative isoform for each gene, an AlphaFold2 pLDDT score track and a predesigned gRNA track. The second table shows a list of introns that satisfy the pLDDT score threshold (Set at < 80 by default, but can be adjusted using the slide bar) along with its intron phase and the pLDDT score in the adjacent 3 amino acids encoded by the surrounding exons. When a row of the second table is selected, up to 3 gRNAs with best gRNA cleavage scores for each intron will be shown in the third table (Figure 4D) and the JBrowse window will zoom into the selected intron. The third table shows the

intron phase, gRNA strand, gRNA sequences and gRNA scores within the selected intron. For convenience, a download link for the csv file containing the information of the selected sRNAs is provided at the bottom of the page (Figure 4D).

Verification of results using public information

For all but two of our target proteins, their localization information of the native protein was made available by The Human Protein Atlas project (Thul et al., 2017), which allowed us to evaluate our knock-in lines. We counterchecked the localization of our target proteins against the Human Protein Atlas data, and 80 out of 87 bulk knock-in lines showed roughly consistent localization results (Figure 5A). For example, our PTBP1 knock-in lines showed nuclear foci, which are known to locate at the perinucleolar compartment (PNC) that are often formed in malignant and transformed cells (Norton and Huang, 2013). Some proteins such as CCT8, PARP1, RACK1, RBM3 and SSRP1 showed localization patterns clearly different from the protein atlas information, but the localization patterns in our library were roughly consistent with the results in other knock-in projects including OpenCell and vpCells (Cho et al., 2022; Reicher et al., 2024). Published immunostaining results of RACK1 and SSRP1 also showed subcellular localization patterns more similar to our results (Adelmant et al., 2012; Chang et al., 2001), suggesting that our tagged proteins for these genes faithfully reported their native localization patterns.

The molecular weights of the 69 target proteins on Western blotting are also good indicator of successful knock-in. 63 out of 69 knock-in lines derived from single cells showed roughly expected molecular sizes (Figure 5B), however, EWSR1, FUS, KHDRBS1, HNRNPU, NCL and SFPQ showed unexpectedly higher sizes. These are consistent with the results shown on the Abcam websites (Anti-EWSR1#ab133288, Anti-FUS#ab154141, Anti-KHDRBS1#ab76471, Anti-HNRNPU#ab172608, Anti-NCL#ab216006, Anti-SFPQ#ab38148), suggesting that the native proteins also show unexpected migration patterns on SDS-PAGE. Indeed, a published paper indicated that the native EWSR1 protein migrated at ~90 kDa unlike its predicted size of 68 kDa, supporting that our Venus-EWSR1 protein was at the correct size (Wang et al., 2022).

The affinity of the Venus tag with various GFP antibodies provides opportunities for biochemical applications. We performed immunoprecipitation

by the anti-GFP VHH antibody using randomly selected 7 knock-in lines derived rom single cells and the target proteins were efficiently purified (Figure 5C, Figure S6). To determine the purity, we stained the gel with the Flamingo dye, and found that the target proteins were purified to homogeneity (Figure 5D). All the 7 immunoprecipitated proteins were detectable by Western blotting (Figure 5C, Figure S6). Therefore, the knock-in resource provides an attractive alternative to immunoprecipitation by antibodies against the native protein, especially when good IP-grade antibodies are not available for the protein of interest, or the user needs to maximize the cost-performance to conduct multiplexed experiments against a large number of target proteins.

Overlaps with the existing database

As parallel efforts, two large-scale knock-in cell lines have been published using HDR (Cho et al., 2022) or intronic NHEJ (Reicher et al., 2024) based on HEK293T cells, with 1,310 and 698 proteins currently targeted, respectively as of February 2025. Although our set currently covers only 87 target proteins, more than half were not present in the vpCells dataset (Figure S7B). Furthermore, although 45 target proteins in our set overlap with those in the OpenCell set, more than a quarter were unique to our set. In addition, we provide a list of 427 gRNAs whose Venus-positivity rates were measured in our project and the availability of individually isolated plasmids that could be used for intronic NHEJ knock-in experiments in different human cell lines using different epitope tags makes our gRNA set a unique resource.

Materials and Methods

Plasmids

To construct the donor plasmid, the Venus sequence was amplified from a plasmid (Gift from the Hiroshi Itoh lab at Nara Institute of Science and Technology) using the Venus F and R primers (Supplementary Table 1). The sequence encoding the linker amino acid sequence of GSGGSGGGS, and the splice acceptor/donor sequence and the gRNA target cleavage sites was synthesized as oligo DNA (Fasmac) and they were combined together with the Venus amplicon and the pUC19 plasmid (Supplementary Table 1) using the In-Fusion HD Cloning Kit (Takara Bio). Three plasmids were produced for different intron phases by using pairs of oligos shown in Supplementary Table 1. In addition, three plasmids were modified to convert Venus to mVenus by using the mVenus mutation F and R primers shown in Supplementary Table 1. The modified plasmids were used in Figure 5A, B, Figure S4 except for mESC and Figure S6. To modify donor DNA cassette with different chemical modifications, the Venus cassette of the plasmid was amplified using Cassette F and R primers (Supplementary Table 1). To obtain the gRNA expression plasmid against target intron shown in Supplementary Table 2, the gRNA sequence predicted by CHOPCHOP (Labun et al., 2019) was synthesized with the sense strand added with 5'-CACC (or 5'-CACCG when the 5' nucleotide was an H) and the antisense strand added with AAAC-3' and inserted in pSpCas9(BB)-2A-miRFP670 (Addgene, #91854) digested with Bbs I (NEB) according to a previous study (Ran et al., 2013). The gRNA sequence against the donor plasmid referred as previous study (Manna et al., 2019; Sawatsubashi et al., 2018; Schmid-Burgk et al., 2016).

Establishment of knock-in cell lines

HEK293T, eHAP1, HeLa, Neuro2a or mESC were grown on 6-well or 24-well plates coated with Cellmatrix Type I-C (Nitta Gelatin) or 0.1% Gelatin and transfected with plasmids or PCR fragments at the ratio of gRNA expression plasmid : Venus donor = 0.076 pmol : 0.064 pmol, using the FuGENE HD Transfection Reagent (Promega) according to the manufacturer's instructions, and FBS (Final concentration 20%) was added to the medium 2 hours after the transfection. MEF or THP-1 cells were transfected with plasmids at the ratio of gRNA expression plasmid : Venus donor = 0.342 pmol : 0.429 pmol, using electroporation using the Neon Transfection System (Life Technologies)

according to the instructions. Cells were resuspended in 1×PBS (-) in the case of detection Venus positivity rates, or Sorting Buffer (1×PBS (-), 1 mM EDTA, 25 mM HEPES-KOH, 1% heat inactivated FBS) in the case of establishment of knock-in cell lines and transferred to 5 mL Round-Bottom Polystyrene Test Tubes with Cell Strainer Cap (Falcon), and Venus positive cells were detected using CytoFLEX S. The cells were sorted onto plates that were coated with Cellmatrix Type I-C (Nitta Gelatin) or 0.1% Gelatin using Cell Sorter MA900 (SONY) or FACS Aria Special Order (BD). After the Venus positive populations were expanded, single Venus positive cells were sorted against onto 96-well plate coated with Cellmatrix Type I-C (Nitta Gelatin) to establish knock-in clones derived from single cells. Acquired FCS files were processed by using flowCore package (Hahne et al., 2009) for visualization on the database.

Organization of representative intron annotation

Matched Annotation from NCBI and EMBL-EBI (MANE v.1.0) (Morales et al., 2022) for human and RefSeq Select (Goldfarb et al., 2025) from (GCF_000001635.27) for mouse were used to determine the representative transcript isoform set. Then, introns in CDSs that intersected with the start or stop codon were excluded from the analysis. The intron phase information was collected for 171,806 introns in 16,993 genes (human) 174,285 introns in 17,715 genes (mouse) in total (Supplementary Table 3). The percentages of intron phase usages were calculated by dividing the number of introns in each phase by the number of total introns in CDSs. The GFF files containing the representative transcript coordinates were generated and sorted by using GFF3sort (Zhu et al., 2017), then, the TBI index files were generated by using Tabix (Li, 2011) for visualization on JBrowse.

Design of gRNA sequences

CHOPCHOP was locally installed under the recommended environment (Python = 2.7.18, Numpy = 1.16.6, Scipy = 1.2.1, pandas = 0.24.02, Biopython = 1.74, argparse = 1.4.0, scikit-learn = 0.18.1)(Labun et al., 2019). gRNAs were designed by CHOPCHOP with the following options: -T 1 -f GN -M NGG -backbone AGGCTAGTCCGT -scoringMethod DOENCH_2016 -G hg38 or mm39 -Target[chr]:Intron[N]$_{start\ site}$ - Intron[N]$_{end\ site}$. The input genomic coordinates were prepared by modifying the annotated intron coordinates. The 75 nt from the exon/intron junction were excluded. In addition, introns longer

than 40,000 nt were split into <=40,000 nt inputs for efficient CHOPCHOP runs. The cleavage efficiencies of gRNAs were predicted based on the machine-learning model developed by Doench et al., (Doench et al., 2016). gRNAs that have potential off-targets with no mismatch were omitted, and those with GC content within the 40-70% range and self-complementarity less than 1 nt were selected. Finally, the three top ranked gRNAs according to the Doench score are shown on the database. In total, this pipeline produced gRNA sequences against human 142,862 introns in 16,362 human genes and 149,561 introns in 17,398 mouse genes (Supplementary Table 4). The Bed files which contain the gRNA coordinates were generated and sorted by using BEDtools (Quinlan and Hall, 2010), then, bed files were converted into BigBed files by using bedToBigBed (Kent and Zahler, 2000) for visualization on JBrowse.

Mapping of pLDDT scores on the human and mouse genomes

PDB files were downloaded from the AlphaFold Protein Structure Database (Varadi et al., 2024), among which 17,485 human proteins and 14,868 mouse proteins matched the MANE v.1.0 or mouse RefSeq Select transcripts. In addition, 15 human proteins (ANXA7, DSP, GSPT1, HNRNPA2B1, HNRNPAB, HNRNPC, HNRNPK, PABPC4, PCBP2, PRRC2C, PTBP1, PTBP3, PUM1, PUM2, XRN1) that from the MANE v.1.0 were individually predicted by AlphaFold2 (Jumper et al., 2021). The predicted pLDDT scores were computationally mapped to the human and mouse genomes at the corresponding codons. In brief, pLDDT scores were extracted by using bio3d package (Grant et al., 2021) and corresponding amino acid coordinates were computationally calculated. Bed files which contain protein coordinate and corresponding pLDDT score were generated and sorted by using BEDtools (Quinlan and Hall, 2010), then, the bed files were converted into BigBed files by using bedToBigBed (Kent and Zahler, 2000) for visualization on JBrowse. The average of pLDDT scores within the 3 amino acids surrounding the predicted tag insertion site was calculated for 139,913 introns in 15,071 human genes and 119,380 introns in 13,119 mouse genes (Supplementary Table 5).

Website

The database was designed using the R shiny package (Chang et al., 2025) and Wix (https://www.wix.com/). In brief, arranged tables with two links were appeared on shiny app, one of link is connected to experimental webpages

designed by Wix homepage editor, other link is connected to subcellular localization information of The Human Protein Atlas (Thul et al., 2017). JBrowse (Diesh et al., 2023) was embedded on the database to visualize the gRNA scores and pLDDT scores on representative transcript isoforms. The shiny app was deployed on shinyApps.io (https://www.shinyapps.io/).

Microscopy

Knock-in cells were grown on 96-well plates (ibidi or MATSUNAMI) coated with 0.1% gelatin in the case of mESCs or Cellmatrix Type I-C (Nitta Gelatin) for other cell lines. Cells were fixed with 4% formaldehyde for 15 minutes, and washed with 0.1% Triton-X for three times followed by a 10-minute permeablization with 0.2% Triton-X and washed with 0.1% Triton-X for three times. Cells were stained with Hoechst33342 for 5 minutes, and observed on ZEISS LSM980 (Carl Zeiss) with a 20× objective lens. The acquired CZI files were processed using ZEN blue edition (Carl Zeiss) and the images were exported to PNG files for visualization on the database.

Splinkerette PCR

Splinkerette PCR was performed according to a protocol modified from a publish method (Potter and Luo, 2010). eHAP1 knock-in cells were pelleted by a centrifugation at 1,200 rpm at 4 degrees for 5 minutes and frozen at -80 degrees. 2 µg of genomic DNA extracted by NucleoSpin Tissue (Takara Bio) was digested by Mse I (NEB) at 37 degrees overnight and the enzymes were heat inactivated at 65 degrees for 20 minutes. The splinkerette adapters (Supplementary Table 6) were mixed and denatured at 95 degrees for 5 minutes, and annealed by cooling the samples at the rate of 0.1 degrees per second until the samples reaches 25 degrees. The annealed linkers were ligated to the digested genomic DNA was purified the FastGene Gel/PCR Extraction Kit (NIPPON Genetics). The target sequence was amplified by nested PCR using KOD FX Neo (TOYOBO) and the primers indicated in Supplementary Table 6. The PCR products were agarose-gel purified using the FastGene Gel/PCR Extraction Kit (NIPPON Genetics) and sequenced by the splinkerette sequencing primer (Supplementary Table 6).

Western blotting

Cells were collected by a centrifugation at 1,200 rpm at 4 degrees for 5 minutes,

and lysed in 2×SDS Lysis Buffer (113.6 mM Tris-HCl pH 6.8, 2.3% SDS, 22.7% Glycerol, 90.9 mM DTT) or Lysis Buffer (50 mM Tris-HCl pH 7.4, 100 mM NaCl, 1% Igepal, 0.1% SDS, 0.5% Sodium deoxycholate). In the case of cell lysis by Lysis Buffer, total protein was quantified with the TaKaRa BCA Protein Assay Kit (Takara) and an appropriate amount of 2×SDS Lysis Buffer was added. After resolving the protein sample by SDS-PAGE, proteins were transferred onto Immobilon-P PVDF Membrane (Merck Millipore) in CAPS Buffer (20% MeOH, 0.05% SDS, 10 mM CAPS pH 11.0), and the membranes were blocked using Blocking Buffer (5% skim milk, 20 mM Tris-HCl pH 7.5, 138 mM NaCl, 0.1% Tween-20) for 30 minutes, and followed by washed with 1×TBST (20 mM Tris-HCl pH 7.5, 138 mM NaCl, 0.1% Tween-20). The membranes were incubated with the primary antibodies listed in Supplementary Table 7 for 1 hour at RT or 1-3 days at 4 degrees and washed five times with 1×TBST followed by incubation with the secondary antibodies (Supplementary Table 7) for 1 hour at room temperature. The membranes were washed again with 1×TBST for five times, and then incubated with ECL Primer Western Blotting Detection Reagent (Cytiva) for 5 minutes. The signals were detected by FUSION-Chemiluminescence Imaging System (Vilber-Lourmat). Acquired gel images were processed for visualization on the database.

Immunoprecipitation

Cells were collected by a centrifugation at 1,200 rpm at 4 degrees for 5 minutes, and lysed in Lysis Buffer containing a Proteinase inhibitor cocktail (Roche or Wako), and incubated the sample at 4 degrees for 5 minutes. The samples were sonicated at Low setting for 5 minutes (30 sec ON, 30 sec OFF, 5 cycles) using Biruptor, and collected supernatant by a centrifugation at 15,000 g at 4 degrees for 10 minutes. Total protein was quantified with the TaKaRa BCA Protein Assay Kit (Takara), and 150-200 µg of total protein was used for immunoprecipitation. 500 µL of the sample diluted with Dilution Buffer (10 mM Tris-HCl pH 7.4, 150 mM NaCl, 0.5 mM EDTA) was rotated with 5 µL of GFP-Trap Magnetic Agarose (Chromo Tek) at 4 degrees for 1 hour. The beads were subsequently washed two times with High-Salt Wash Buffer (50 mM Tris-HCl pH 7.4, 1M NaCl, 1 mM EDTA, 1% Igepal, 0.1% SDS, 0.5% Sodium deoxycholate), 1 time with High-Salt Wash Buffer + Wash Buffer (20 mM Tris-HCl pH 7.4, 10 mM $MgCl_2$, 5 mM NaCl, 0.2% Tween-20), 3 times with Wash Buffer. The washed beads were resuspended in an appropriate amount of 2×SDS Lysis Buffer, and the mixtures

were incubated at 98 degrees for 5 minutes. The supernatant was separated from beads on a magnetic tube rack. After resolving the protein sample by SDS-PAGE, gels were fixed with Fix solution (40% EtOH, 10% Acetic Acid) for 2 hours, followed by stained with Flamingo Fluorescent Protein Gel Stain (Bio-Rad). Western blotting was done as mentioned above.

Data availability

The entry, gRNA database and collection pages are available at <https://yumahanaiatokamuralab.shinyapps.io/KnockInAtlas/>, <https://yumahanaiatokamuralab.shinyapps.io/KnockInAtlas-Human-gRNA/>, <https://yumahanaiatokamuralab.shinyapps.io/KnockInAtlas-Mouse-gRNA/> and <https://yumahanaiatokamuralab.shinyapps.io/KnockInAtlas-Collection/>, respectively.

Conflict of interest

The authors declare that they have no conflicts of interest with the contents of this article.

Acknowledgements

The authors thank Yuji Shimizu for mESC culture, Hiroshi Ito for the Venus plasmid and the Life Science Collaboration Center (LiSCo) facilities for microscopy. We also thank Masami Shiimori, Ayako Isotani, Toshiaki Shigeoka, Yasumasa Ishida, Minato Hirano and Kentaro Yoshii for constructive feedback on the manuscript and the website. We also thank Yae Toda and Junko Iida for technical assistance.

Author contributions

Y.H., P.L.L.H., Y.S., and K.O. writing-review and editing; Y.H. and K.O. writing-original draft; Y.H. visualization; Y.H. software; Y.H., P.L.L.H., Y.S., N.K., M.K. and K.M. resources; N.K., T.K., Y.H., P.L.L.H., Y.S., and K.O. methodology; Y.H. and K.O. investigation; N.K., T.K., Y.H. and K.O. formal analysis; N.K., T.K., Y.H. and K.O. funding acquisition; T. K. and K.O. supervision; T.K. and K.O. project administration; K.O. conceptualization.

Funding and additional information

Y.H. was supported by the research fellowship from the NAIST for the Creation of Innovation in Science and Technology, and the JSPS Research Fellowship for Young Scientists (Grant-in-Aid for JSPS Fellows, 24KJ1692). Works at K.O.'s lab was supported by the JSPS Found for the Promotion of Joint International Research (Home-Returning Research Development Research, 17K20145), Academic Assistant Grant by Office for Gender Equality at NAIST and the Takeda Science Foundation. T.K.'s lab was supported by JSPS KAKENHI (Grant-in-Aid for Scientific Research (B) 20H03468). N.K. was supported by JSPS KAKENHI (Grant-in-Aid for Early-Career Scientists 23K14546).

References

Adelmant, G., Calkins, A.S., Garg, B.K., Card, J.D., Askenazi, M., Miron, A., Sobhian, B., Zhang, Y., Nakatani, Y., Silver, P.A., et al. (2012). DNA Ends Alter the Molecular Composition and Localization of Ku Multicomponent Complexes. Mol. Cell. Proteomics *11*, 411–421.

Banan, M. (2020). Recent advances in CRISPR/Cas9-mediated knock-ins in mammalian cells. J. Biotechnol. *308*, 1–9.

Binder, J.L., Berendzen, J., Stevens, A.O., He, Y., Wang, J., Dokholyan, N. V., and Oprea, T.I. (2022). AlphaFold illuminates half of the dark human proteins. Curr. Opin. Struct. Biol. *74*, 102372.

Chang, B.Y., Chiang, M., and Cartwright, C.A. (2001). The Interaction of Src and RACK1 Is Enhanced by Activation of Protein Kinase C and Tyrosine Phosphorylation of RACK1. J. Biol. Chem. *276*, 20346–20356.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., A, D., and B, B. (2025). shiny: Web Application Framework for R. R package version 1.10.0.9000.

Cho, N.H., Cheveralls, K.C., Brunner, A.-D., Kim, K., Michaelis, A.C., Raghavan, P., Kobayashi, H., Savy, L., Li, J.Y., Canaj, H., et al. (2022). OpenCell: Endogenous tagging for the cartography of human cellular organization. Science (80-.). *375*.

Diesh, C., Stevens, G.J., Xie, P., De Jesus Martinez, T., Hershberg, E.A., Leung, A., Guo, E., Dider, S., Zhang, J., Bridge, C., et al. (2023). JBrowse 2: a modular genome browser with views of synteny and structural variation. Genome Biol. *24*, 74.

Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., et al. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. *34*, 184–191.

George, N., Fexova, S., Fuentes, A.M., Madrigal, P., Bi, Y., Iqbal, H., Kumbham, U., Nolte, N.F., Zhao, L., Thanki, A.S., et al. (2024). Expression Atlas update: insights from sequencing data at both bulk and single cell level. Nucleic Acids Res. *52*, D107–D114.

Goldfarb, T., Kodali, V.K., Pujar, S., Brover, V., Robbertse, B., Farrell, C.M., Oh, D.-H., Astashyn, A., Ermolaeva, O., Haddad, D., et al. (2025). NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. Nucleic Acids Res. *53*, D243–D257.

Grant, B.J., Skjærven, L., and Yao, X. (2021). The <scp>Bio3D</scp> packages for structural bioinformatics. Protein Sci. *30*, 20–30.

Hahne, F., LeMeur, N., Brinkman, R.R., Ellis, B., Haaland, P., Sarkar, D., Spidlen, J., Strain, E., and Gentleman, R. (2009). flowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinformatics *10*, 106.

Jin, H., Zhang, C., Zwahlen, M., von Feilitzen, K., Karlsson, M., Shi, M., Yuan, M., Song, X., Li, X., Yang, H., et al. (2023). Systematic transcriptional analysis of human cell lines for gene expression landscape and tumor representation. Nat. Commun. *14*, 5417.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

Kent, W.J., and Zahler, A.M. (2000). The intronerator: exploring introns and alternative splicing in Caenorhabditis elegans. Nucleic Acids Res *28*, 91–93.

Labun, K., Montague, T.G., Krause, M., Torres Cleuren, Y.N., Tjeldnes, H., and Valen, E. (2019). CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. Nucleic Acids Res. *47*, W171–W174.

Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics *27*, 718–719.

Lienert, F., Mohn, F., Tiwari, V.K., Baubec, T., Roloff, T.C., Gaidatzis, D., Stadler, M.B., and Schübeler, D. (2011). Genomic Prevalence of Heterochromatic H3K9me2 and Transcription Do Not Discriminate Pluripotent from Terminally Differentiated Cells. PLoS Genet. *7*, e1002090.

Manna, P.T., Davis, L.J., and Robinson, M.S. (2019). Fast and cloning□free CRISPR/Cas9□mediated genomic editing in mammalian cells. Traffic *20*, 974–982.

Morales, J., Pujar, S., Loveland, J.E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C.M., et al. (2022). A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature *604*, 310–315.

Norton, J.T., and Huang, S. (2013). The Perinucleolar Compartment: RNA Metabolism and Cancer. pp. 139–152.

Potter, C.J., and Luo, L. (2010). Splinkerette PCR for Mapping Transposable Elements in Drosophila. PLoS One *5*, e10168.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. *8*, 2281–2308.

Reicher, A., Koren, A., and Kubicek, S. (2020). Pooled protein tagging, cellular imaging, and in situ sequencing for monitoring drug action in real time. Genome Res. *30*, 1846–1855.

Reicher, A., Reiniš, J., Ciobanu, M., Růžička, P., Malik, M., Siklos, M., Kartysh, V., Tomek, T., Koren, A., Rendeiro, A.F., et al. (2024). Pooled multicolour tagging for visualizing subcellular protein dynamics. Nat. Cell Biol. *26*, 745–756.

Sawatsubashi, S., Joko, Y., Fukumoto, S., Matsumoto, T., and Sugano, S.S. (2018). Development of versatile non-homologous end joining-based knock-in module for genome editing. Sci. Rep. *8*, 593.

Schmid-Burgk, J.L., Höning, K., Ebert, T.S., and Hornung, V. (2016). CRISPaint allows modular base-specific gene tagging using a ligase-4-dependent mechanism. Nat. Commun. *7*, 12338.

Serebrenik, Y. V., Sansbury, S.E., Kumar, S.S., Henao-Mejia, J., and Shalem, O. (2019). Efficient and flexible tagging of endogenous genes by homology-independent intron targeting. Genome Res. *29*, 1322–1328.

Thul, P.J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L.M., et al. (2017). A subcellular map of the human proteome. Science (80-.). *356*.

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., et al. (2024). AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res. *52*, D368–D375.

Wachutka, L., Caizzi, L., Gagneur, J., and Cramer, P. (2019). Global donor and acceptor splicing site kinetics in human cells. Elife *8*.

Wang, Y., Feswick, A., Apostolou, V., Petkov, P.M., Moser, E.K., and Tibbetts, S.A. (2022). Gammaherpesvirus-mediated repression reveals EWSR1 to be a negative regulator of B cell responses. Proc. Natl. Acad. Sci. *119*.

Yu, Y., Guo, Y., Tian, Q., Lan, Y., Yeh, H., Zhang, M., Tasan, I., Jain, S., and Zhao, H. (2020). An efficient gene knock-in strategy using 5′-modified double-stranded DNA donors with short homology arms. Nat. Chem. Biol. *16*, 387–390.

Zhong, H., Ceballos, C.C., Massengill, C.I., Muniak, M.A., Ma, L., Qin, M., Petrie, S.K., and Mao, T. (2021). High-fidelity, efficient, and reversible labeling of

endogenous proteins using CRISPR-based designer exon insertion. Elife *10*.
Zhu, T., Liang, C., Meng, Z., Guo, S., and Zhang, R. (2017). GFF3sort: a novel tool to sort GFF3 files for tabix indexing. BMC Bioinformatics *18*, 482.

Figure legends

Figure 1 Evaluation of donor DNA in different forms

(A) Schematic of the strategy of the plasmid-based NHEJ knock-in. Three plasmids are simultaneously transfected: 1. gRNA expression plasmid to excise the Venus cassette from the donor plasmid, 2. gRNA expression plasmid to cleave the target intron, 3. Donor plasmid containing the Venus sequence flanked by the splice acceptor (SA) and splice donor (SD). Depending on the reading frame of the surrounding exons one of the three donor plasmids with the correct reading frame (Intron 0, 1 or 2) is used.

(B) Utilization of modified PCR products in NHEJ knock-in. Instead of relying on the excision of the Venus cassette from the plasmid by CRISPR/Cas9 in cells, knock-in using donor DNA in the form of PCR products with different chemical modification at the 5' end was examined. In addition to PCR products produced with regular 5'-hydroxyl DNA primers, those with 2PS (phosphorothioate linkages at the first two nucleotides), 5'-C6, 5'-C6-PEG10 or 5'-C6-Biotin modification were examined.

(C) HNRNPA1 gene locus. A gRNA targeting HNRNPA1-Intron1 was used in experiments in panels (D) and (E).

(D) Flow cytometry results of knock-in cells 2 days after transfection. The y-axis shows the percentage of Venus positive cells within the transfected cell population as judged by the presence of miRFP670 fluorescence derived from the gRNA expression plasmid backbone. The scatter plots used to calculate Venus positivity rates are shown in Figure S2A. Grey bars show the knock-in results with the non-targeting gRNA expression plasmid, gRNA expression plasmid against the donor sequence and the Venus donor (in the case of plasmids) or the PCR product. Green bars show the results with the complete set of gRNA expression plasmids and donor DNA. Data represent mean±SD (n = 3).

(E) Flow cytometry results of knock-in cells at the second sorting. Cells after initial Venus positive sorting shown in green bars in (D) were grown, and the resulting populations were analyzed again by flow cytometry. The scatter plots used to calculate Venus positivity rates are shown in Figure S2B. Data represent mean±SD

Figure 2 gRNA design pipeline

(A) Schematic of the gRNA prediction strategy. The representative transcripts were taken from the MANE v.1.0 and Refseq Select references, and used the intron sequences for the CHOPCHOP prediction to obtain gRNA sequences. The parameters used are shown.

(B) Visualization of pLDDT and gRNA scores on UCSC genome browser. The pLDDT scores were extracted from the AlphaFold Structure database by the bio3d package, and the scores were mapped to the genome at the corresponding codons. The pLDDT scores are color coded on the pLDDT score track, and the mapped scores at the human HNRNPA1 locus is shown as an example.

(C) Distribution of mean of pLDDT scores within 3 amino acids surrounding the predicted tag insertion site. The averages of the pLDDT scores for the surrounding 3 amino acids were calculated for all human and mouse introns in the representative transcript set, and the histograms for human and mouse were drawn. The x and y-axes show pLDDT score and the number of introns in each bin, respectively.

(D) Correlation analysis between RNA expression levels in HEK293 cells and Venus positivity rates against each target gene. The observed Venus median fluorescent intensity (MFI) after transfection were plotted against the mRNA expression level of the corresponding gene in HEK293 cells in a published RNAseq data (Jin et al., 2023). The correlation coefficient and p-values were calculated by spearman's correlation analysis (n = 213).

Figure 3 Experimental summary of HEK293T knock-in lines

(A) Screenshot of the experimental summary page in the HEK293T knock-in line database. The “HEK293T knock-in cell lines” page provides a table with last two columns showing URLs, one of which prompts the webpage summarizing our knock-in results (magenta rectangle) and the other the webpage of The Human Protein Atlas (dark blue rectangle).

(B) The webpage of experimental summary for each target intron. The webpage contains 3 panes (i) Flow cytometry data of Venus positivity rates 2 days after transfection (ii) Fluorescent images of each bulk knock-in line (knock-in lines derived from single cells in the case of HNRNPA1) (iii) Western blots of each knock-in line derived from single cells.

Figure 4 gRNA database for human and mouse

Screenshots of the gRNA database. Users can use the database to look for gRNAs for a target of interest using this database. There “Human gRNA database” and “Mouse gRNA database” pages and the human database is shown as an example.

(A) The first table shows a list of genes, genomic coordinates and representative gene functions.

(B) The second table shows a list of introns corresponding to the selected gene to show the intron phase, intron coordinates and pLDDT score.

(C) The JBrowse window shows the selected gene locus or intron with three tracks, which display the representative transcript isoform, pLDDT score and gRNA locations.

(D) The third table shows a list of gRNA sequences corresponding to the selected intron along with the information of the intron phase, gRNA spacer sequence, gRNA PAM sequence and gRNA score.

Figure 5 Examples of experiments using HEK293T knock-in lines

(A) Representative confocal microscope images of tagged proteins. Venus positive cells were enriched by a cell sorter and stained with Hoechst. The pictures were taken with a 20× objective lens on a confocal microscope, and scale bars are 50 μm.

(B) Representative western blots of knock-in lines derived from single cells. Lysates from the indicated knock-in lines were resolved by 8% SDS-PAGE and bands were detected by an anti-GFP and anti-GAPDH antibodies sequentially.

(C) Representative western blot of Venus-tagged HNRNPA1 protein after immunoprecipitation using anti-GFP VHH antibody. The HNRNPA1-Intron1 line derived from single cells was used and the protein was immunoprecipitated with the anti-GFP VHH antibody. Samples were resolved 12% SDS-PAGE and bands were detected by an anti-GFP and anti-β-Tubulin antibodies sequentially.

(D) Purification of the Venus-tagged HNRNPA1 protein. Parental HEK293T was used as a negative control. The 12% SDS-PAGE gel was stained with flamingo staining solution and bands were detected by a fluorescence gel imager.

Figure S1 Percentages of intron phases in CDSs in the human and mouse genomes

The pie chart shows the fraction of Intron (0), (1) and (2) among (A) the 171,806 introns in the human 16,993 genes and (B) the 174,285 introns in the mouse 17,715 genes.

Figure S2 Fluorescence intensity of cells after Venus knock-in at HNRNPA1-Intron 1

Scatter plots of log10-transformed Venus intensity (x-axis) and FSC-A (y-axis) are shown. Flow cytometry analysis was done 2 days after transfection (A) or 1 week after sorting (B).

Figure S3 Venus positivity rates by various donors

(A) Flow cytometry data of knock-in cells 2 days after transfection. Grey bars show the results with the non-targeting gRNA expression plasmid, gRNA expression plasmid against the donor sequence and the Venus donor (in the case of plasmids) or the PCR product. Green bars show the results with the complete set gRNA expression plasmid and donor DNA.

(B) Flow cytometry data of knock-in THP-1 and MEF cells 1-2 weeks after transfection. Scatter plots of log10-transformed Venus intensity (x-axis) and FCS-A (y-axis) are shown.

Figure S4 Subcellular localization of HNRNPA1-Venus

The fusion protein localization of HNRNPA1-Venus was observed in the bulk knock-in cell lines based on (A) HEK293T, (B) eHAP1, (C) HeLa, (D) Neuro2a and (E) mESC with a 20× objective lens, and scale bars correspond to 50 μm. PCR product with 5'-C6 modification were used except for mESC (plasmid donor was used in the case of mESC).

Figure S5 Determination of insertion sites using splinkerette PCR

(A) Schematic of the splinkerette PCR method. Genomic DNA digested with Mse I is ligated with splinkerette adapter and DNA fragments were amplified using Venus-specific primer and splinkerette adapter-specific primer.

(B) The sequences and the predicted structure of the Splinkerette adapter. The upper strand has the TA sticky 5' overhang (underline).

(C) Representative agarose gel picture of splinkerette PCR using the HNRNPA1

knock-in eHAP1 clones derived from single cells. Most clones produced single bands at the expected around 700 bp size, except clone 5 that gave rise to two bands.

Figure S6 Summary of immunoprecipitations

Western blot of mVenus-tagged protein after immunoprecipitation using anti-GFP VHH antibody. The FUS-Intron10, DDX3X-Intron1, G3BP1-Intron5, G3BP2-Intron5, HNRNPM-Intron9 and SYNCRIP-Intron4 lines derived from single cells were immunoprecipitated with the anti-GFP VHH antibody. Samples were resolved 8% SDS-PAGE and bands were detected by an anti-GFP and anti-β-Tubulin or anti-GAPDH antibodies sequentially.

Figure S7 Summary of our knock-in library

(A) Venus positivity rates of knock-in cells for 177 gens 425 introns. Flow cytometry of knock-in cells 2 days after transfection against 425 introns (intron-0: n = 183, intron-1: n = 148, intron-2: n = 94). The y-axis shows the percentage of Venus positive cells within the transfected cell population as judged by the presence of miRFP670 fluorescence derived from the gRNA expression plasmid backbone. 213 out of 425 shows more than 0.1% of Venus positivity rates. Data represent mean (n = 3) except for EIF3E-Intron8 (n = 2).

(B) Overlaps with other knock-in resources. The Ven diagram shows the overlaps of HEK293T knock-in cell lines between our project (n = 87 genes), OpenCell (n = 1,310) and vpCells (n = 698).

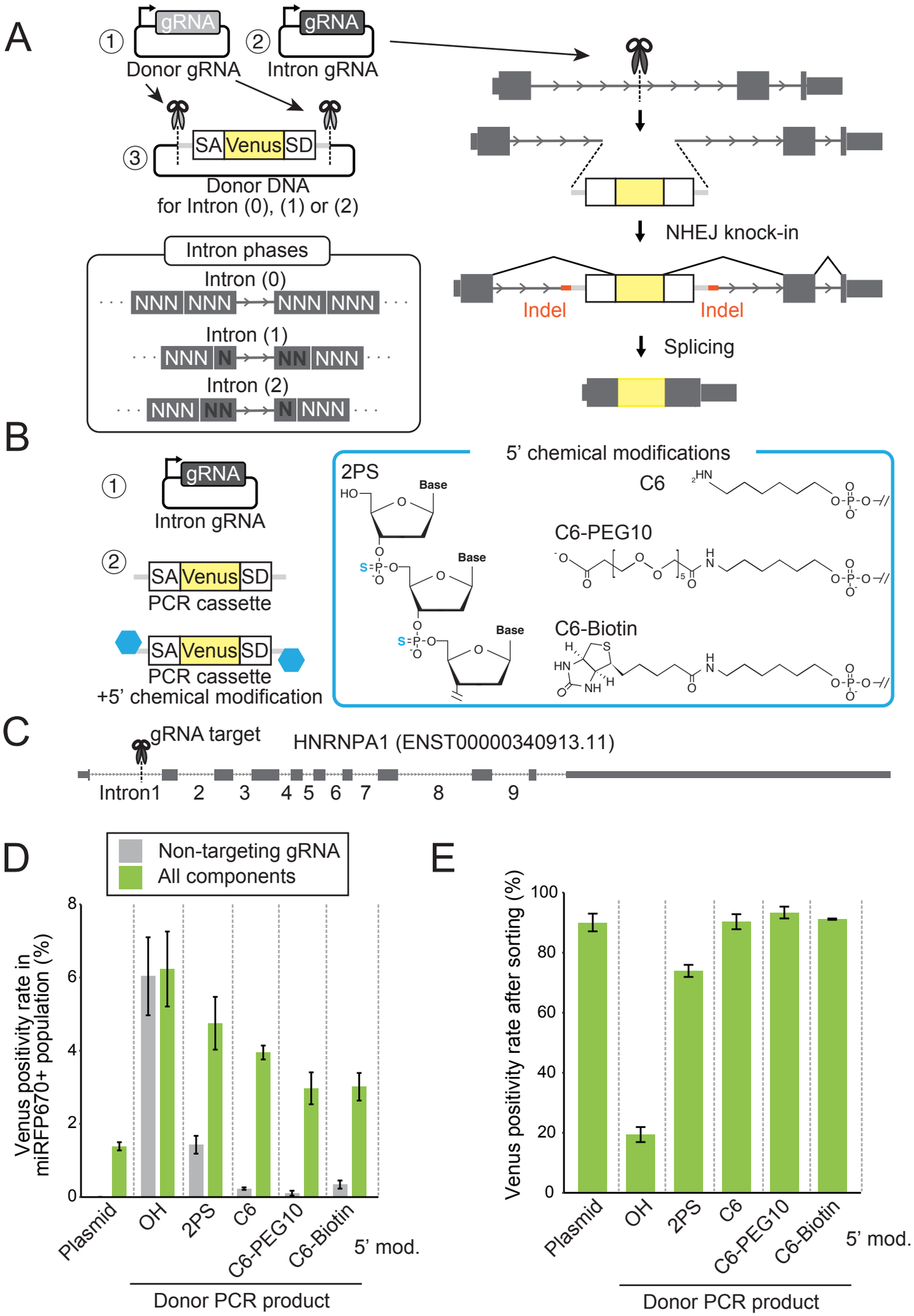

A
gRNA
Donor gRNA
Intron gRNA
SA Venus SD
Donor DNA
for Intron (0), (1) or (2)
Intron phases
Intron (0)
NNN NNN NNN NNN
Intron (1)
NNN N NN NNN
Intron (2)
NNN NN N NNN
NHEJ knock-in
Indel
Indel
Splicing
B
gRNA
Intron gRNA
SA Venus SD
PCR cassette
SA Venus SD
PCR cassette
+5' chemical modification
5' chemical modifications
2PS
Base
C6
C6-PEG10
C6-Biotin
C
gRNA target
HNRNPA1 (ENST00000340913.11)
Intron1 2 3 4 5 6 7 8 9
D
Non-targeting gRNA
All components
Venus positivity rate in miRFP670+ population (%)
Plasmid
OH
2PS
C6
C6-PEG10
C6-Biotin
5' mod.
Donor PCR product
E
Venus positivity rate after sorting (%)
Plasmid
OH
2PS
C6
C6-PEG10
C6-Biotin
5' mod.
Donor PCR product

Figure 1

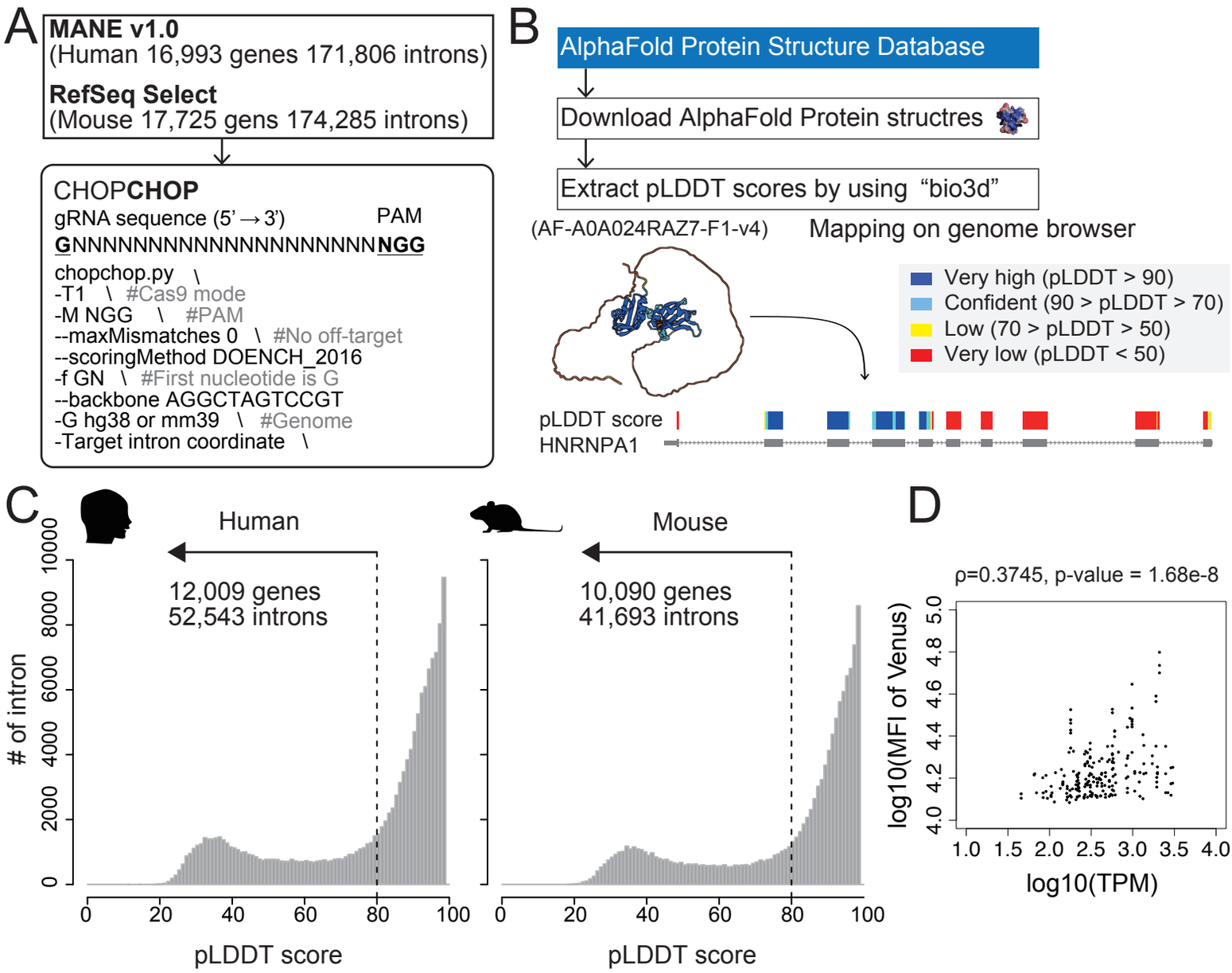

A
MANE v1.0
(Human 16,993 genes 171,806 introns)
RefSeq Select
(Mouse 17,725 gens 174,285 introns)
CHOPCHOP
gRNA sequence (5' → 3')
PAM
GNNNNNNNNNNNNNNNNNNNNNNGG
chopchop.py \
-T1 \ #Cas9 mode
-M NGG \ #PAM
--maxMismatches 0 \ #No off-target
--scoringMethod DOENCH_2016
-f GN \ #First nucleotide is G
--backbone AGGCTAGTCCGT
-G hg38 or mm39 \ #Genome
-Target intron coordinate \
B
AlphaFold Protein Structure Database
Download AlphaFold Protein structres
Extract pLDDT scores by using "bio3d"
(AF-A0A024RAZ7-F1-v4)
Mapping on genome browser
Very high (pLDDT > 90)
Confident (90 > pLDDT > 70)
Low (70 > pLDDT > 50)
Very low (pLDDT < 50)
pLDDT score
HNRNPA1
C
Human
12,009 genes
52,543 introns
Mouse
10,090 genes
41,693 introns
# of intron
pLDDT score
D
ρ=0.3745, p-value = 1.68e-8
log10(MFI of Venus)
log10(TPM)

Figure 2

A

THE KNOCK-IN ATLAS

HEK293T knock-in cell lines

HEK293T knock-in cell lines | Instruction | Links | Contact us

Show 10 entries

Search:

Transcript_ID	Gene	Gene_intron	Intron_phase	FACS	Microscope	WB	Link	ProteinAtlasLink
ENST00000340913.11	HNRNPA1	HNRNPA1_In8	Intron(1)	O	O	O	Knock-in site (HNRNPA1_In8)	ProteinAtlas (HNRNPA1)
ENST00000618183.5	HNRNPA2B1	HNRNPA2B1_In3	Intron(0)	O	O	O	Knock-in site (HNRNPA2B1_In3)	ProteinAtlas (HNRNPA2B1)
ENST00000618183.5	HNRNPA2B1	HNRNPA2B1_In6	Intron(1)	O	O	O	Knock-in site (HNRNPA2B1_In6)	ProteinAtlas (HNRNPA2B1)
ENST00000392524.7	HNRNPA3	HNRNPA3_In7	Intron(1)	O	O	O	Knock-in site (HNRNPA3_In7)	ProteinAtlas (HNRNPA3)
ENST00000358344.8	HNRNPAB	HNRNPAB_In6	Intron(1)	O	O	O	Knock-in site (HNRNPAB_In6)	ProteinAtlas (HNRNPAB)
ENST00000553300.6	HNRNPC	HNRNPC_In6	Intron(1)	O	O	O	Knock-in site (HNRNPC_In6)	ProteinAtlas (HNRNPC)
ENST00000313899.12	HNRNPD	HNRNPD_In2	Intron(2)	O	O	O	Knock-in site (HNRNPD_In2)	ProteinAtlas (HNRNPD)
ENST00000295470.10	HNRNPDL	HNRNPDL_In1	Intron(2)	O	O	O	Knock-in site (HNRNPDL_In1)	ProteinAtlas (HNRNPDL)
ENST00000393432.9	HNRNPH1	HNRNPH1_In2	Intron(1)	O	O	O	Knock-in site (HNRNPH1_In2)	ProteinAtlas (HNRNPH1)
ENST00000265866.12	HNRNPH3	HNRNPH3_In7	Intron(1)	O	O	O	Knock-in site (HNRNPH3_In7)	ProteinAtlas (HNRNPH3)

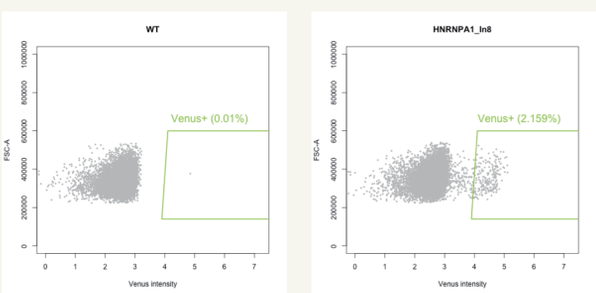
Showing 1 to 10 of 427 entries

Previous 1 2 3 4 5 ... 43 Next

B

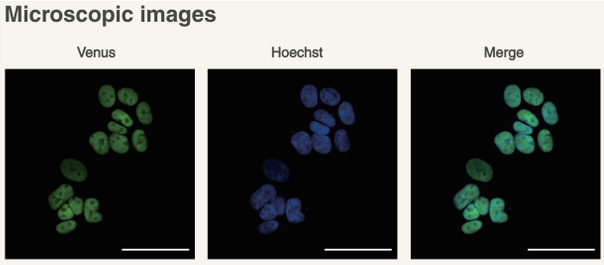
(i)

Flow cytometry data



(ii)

Microscopic images



Objective: ×20, Scale bar: 50 µm, Scan zoom: ×3

(iii)

Western blots

Expected fusion protein MW (kDa): 66.8

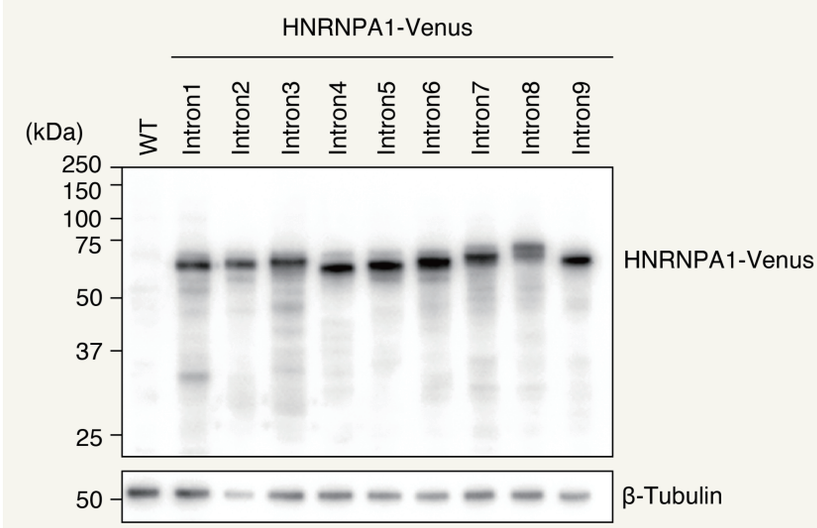


Figure 3

A

THE KNOCK-IN ATLAS

Human Intron gRNA Databse

Human gRNA Database | Instruction | Download | Links | Contact us

Show 10 entries

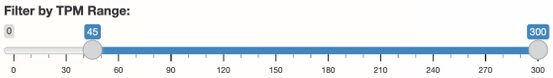
Search:

Gene	chromosome	start	end	HEK293_TPM	HeLa_TPM	HAP1_TPM	THP1_TPM	Function
HNRNPA1	chr12	54280726	54287087	980.1	1159.3	1366.5	1290.2	mRNA processing, Transport, Host-virus interaction
HNRNPA2B1	chr7	26189927	26200746	179.6	273.7	180.1	294.9	mRNA processing, Transport, Host-virus interaction
HNRNPD	chr4	82352498	82373991	424.8	409.8	683.8	520.1	Biological rhythms, Transcription
HNRNPM	chr19	8444975	8489114	470	398.8	553.6	437.4	mRNA processing
HNRNPK	chr9	83968083	83980215	568.6	679.9	990.2	728.6	mRNA processing, Transcription, Host-virus interaction
HNRNPU	chr1	244850297	244864543	267.3	260.7	387.5	410	Biological rhythms, Cell cycle, Differentiation, mRNA processing, Transcription, Host-virus interaction
FUS	chr16	31180139	31191605	286.1	376.4	517.2	580.3	
G3BP1	chr5	151771954	151812785	163.3	189.3	293.4	144.3	Immunity, Transport, Host-virus interaction
G3BP2	chr4	75642786	75673411	62.1	104.9	156	92.1	Immunity, Transport, Host-virus interaction
SAMD11	chr1	923923	944574	82.1	98.1	8.1	0.8	Transcription

Showing 1 to 10 of 2,516 entries

Previous 1 2 3 4 5 ... 252 Next

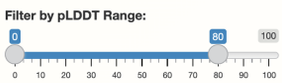
Filter by TPM Range:



B

gRNAs for the selected gene:

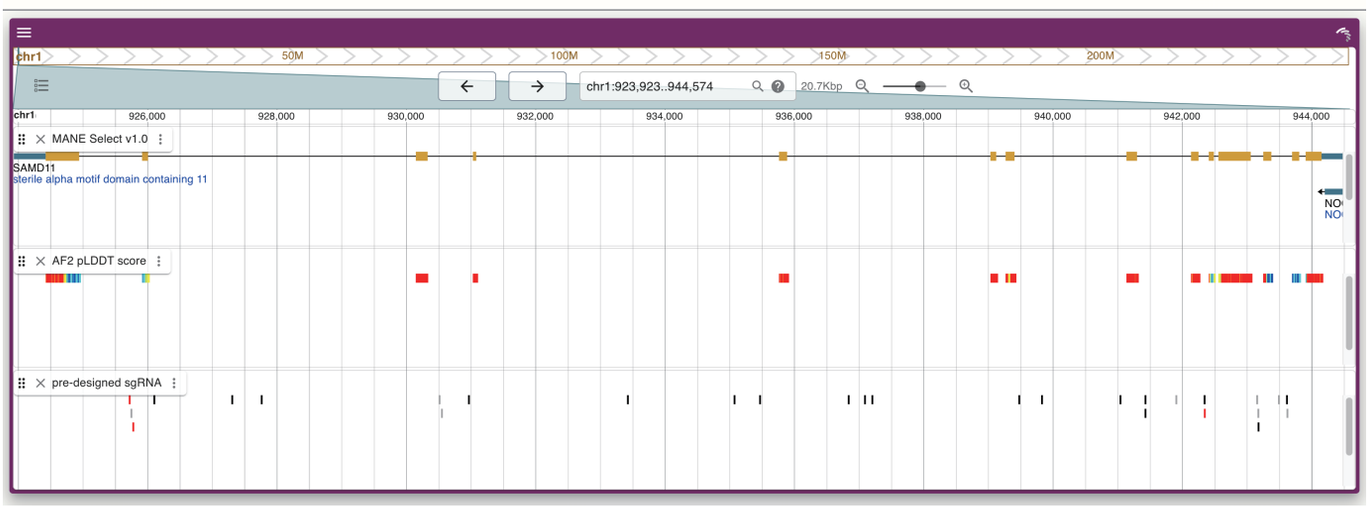
Filter by pLDDT Range:



Gene	IntronID	IntronPhase	chromosome	start	end	pLDDT
SAMD11	SAMD11_In2	Intron(0)	chr1	926014	930154	49.48
SAMD11	SAMD11_In3	Intron(2)	chr1	930337	931038	30.86
SAMD11	SAMD11_In4	Intron(2)	chr1	931090	935771	32.38
SAMD11	SAMD11_In5	Intron(1)	chr1	935897	939039	31.03
SAMD11	SAMD11_In6	Intron(1)	chr1	939130	939274	26.89
SAMD11	SAMD11_In7	Intron(1)	chr1	939413	941143	39.74
SAMD11	SAMD11_In8	Intron(2)	chr1	941307	942135	29.48
SAMD11	SAMD11_In9	Intron(1)	chr1	942252	942409	43.08
SAMD11	SAMD11_In10	Intron(2)	chr1	942489	942558	69.79
SAMD11	SAMD11_In11	Intron(1)	chr1	943059	943252	28.31

Genome Web Browser

C



D

gRNAs for the selected intron:

Gene	IntronID	IntronPhase	gRNAID	gRNA_strand	gRNA_spacer	gRNA_PAM	gRNA_score
SAMD11	SAMD11_In5	Intron(1)	SAMD11_In5_TOP1	-	GCGTATGGGAGACCTGGACA	CGG	71.4
SAMD11	SAMD11_In5	Intron(1)	SAMD11_In5_TOP2	+	GCCCCATCGATCCAGCAGAG	CGG	69.21
SAMD11	SAMD11_In5	Intron(1)	SAMD11_In5_TOP3	+	GGGAGACGATGGCTTCTCAG	GGG	67.14

Download selected gRNA

Figure 4

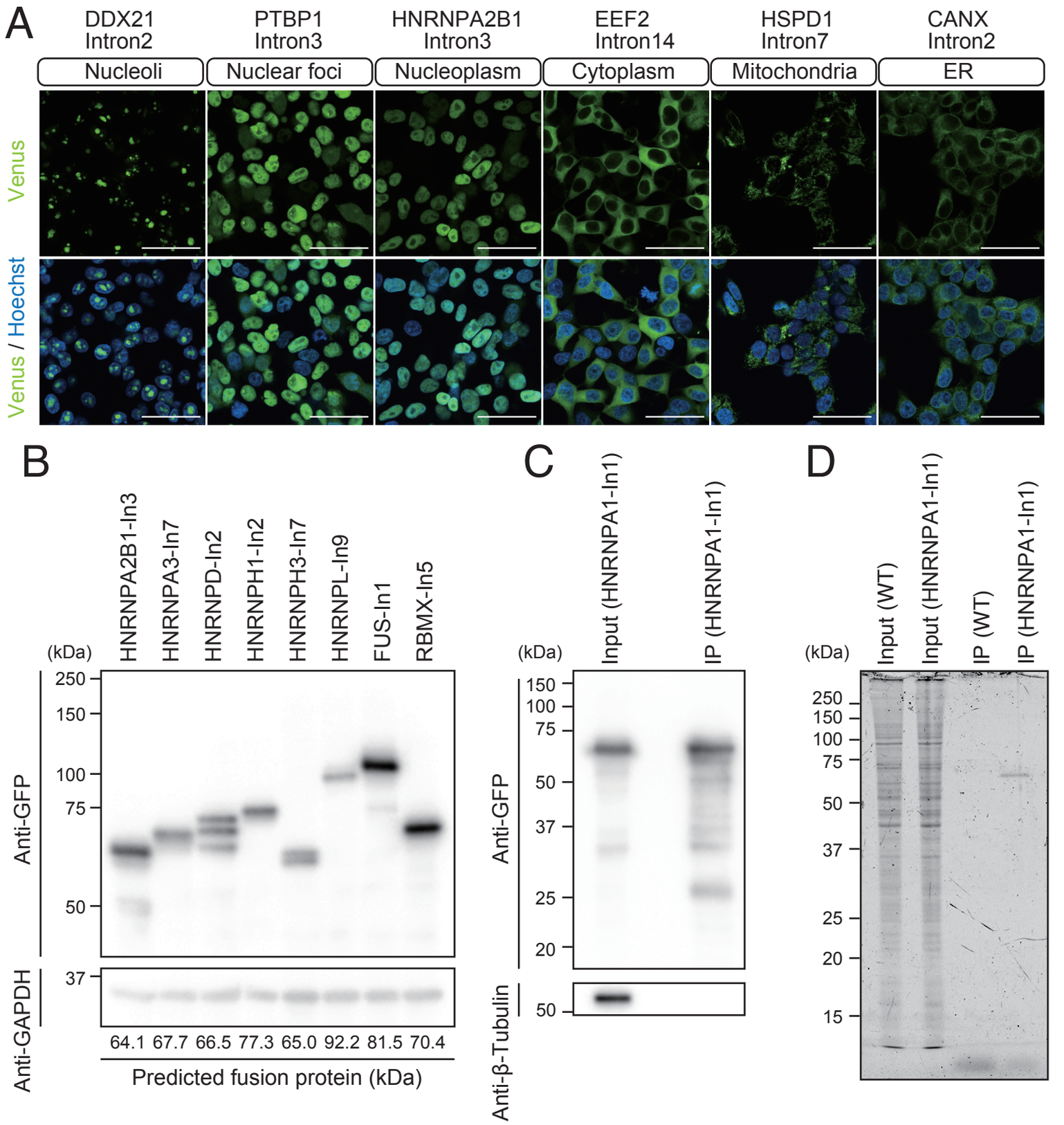
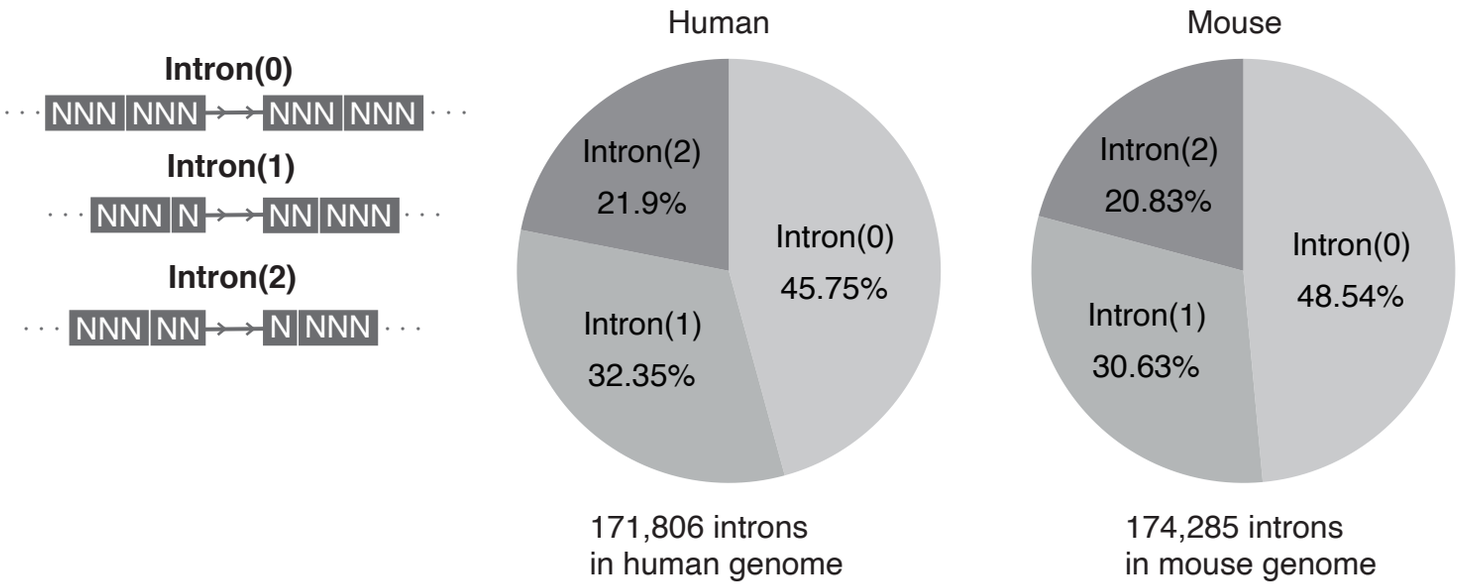
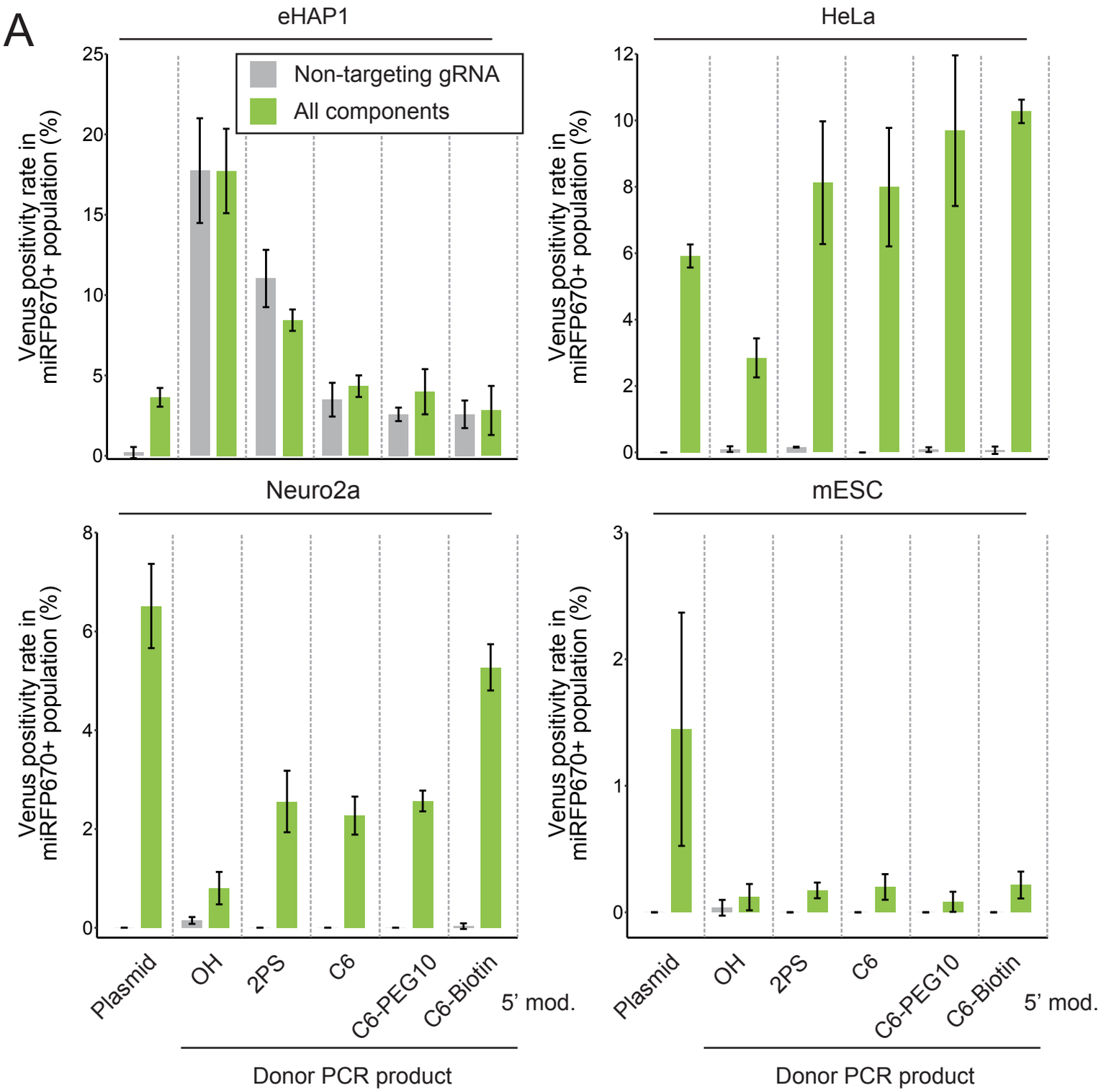

A
DDX21 Intron2
PTBP1 Intron3
HNRNPA2B1 Intron3
EEF2 Intron14
HSPD1 Intron7
CANX Intron2
Nucleoli
Nuclear foci
Nucleoplasm
Cytoplasm
Mitochondria
ER
Venus
Venus / Hoechst
B
HNRNPA2B1-In3
HNRNPA3-In7
HNRNPD-In2
HNRNPH1-In2
HNRNPH3-In7
HNRNPL-In9
FUS-In1
RBMX-In5
(kDa)
250
150
100
75
50
37
Anti-GFP
Anti-GAPDH
64.1 67.7 66.5 77.3 65.0 92.2 81.5 70.4
Predicted fusion protein (kDa)
C
Input (HNRNPA1-In1)
IP (HNRNPA1-In1)
(kDa)
150
100
75
50
37
25
20
Anti-GFP
Anti-β-Tubulin
D
Input (WT)
Input (HNRNPA1-In1)
IP (WT)
IP (HNRNPA1-In1)
(kDa)
250
150
100
75
50
37
25
20
15

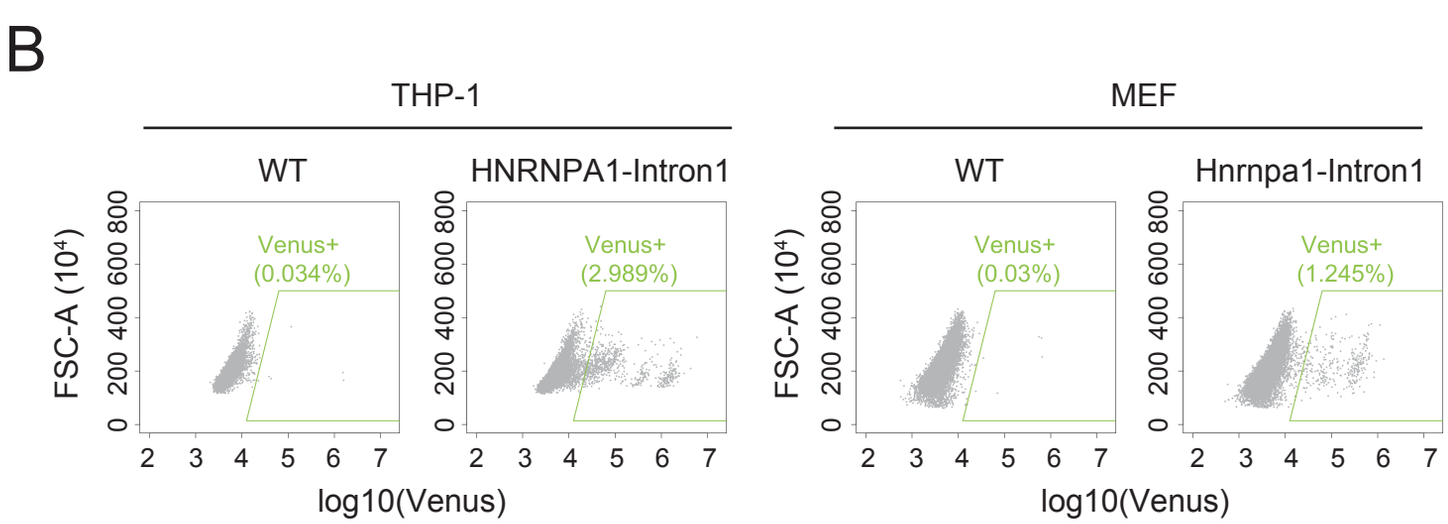
Figure 5



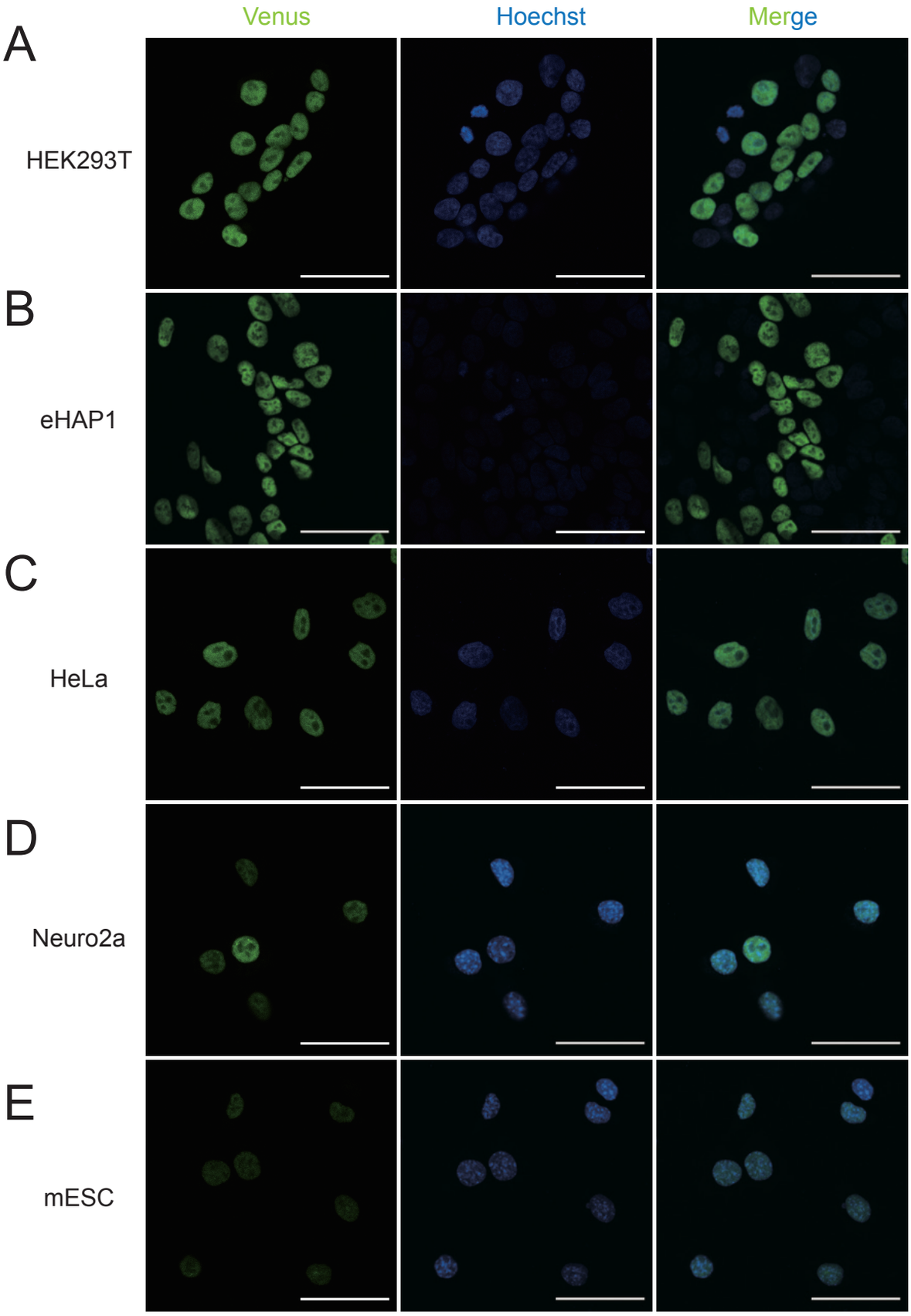
Supplementary Figure 1

Supplementary Figure 2

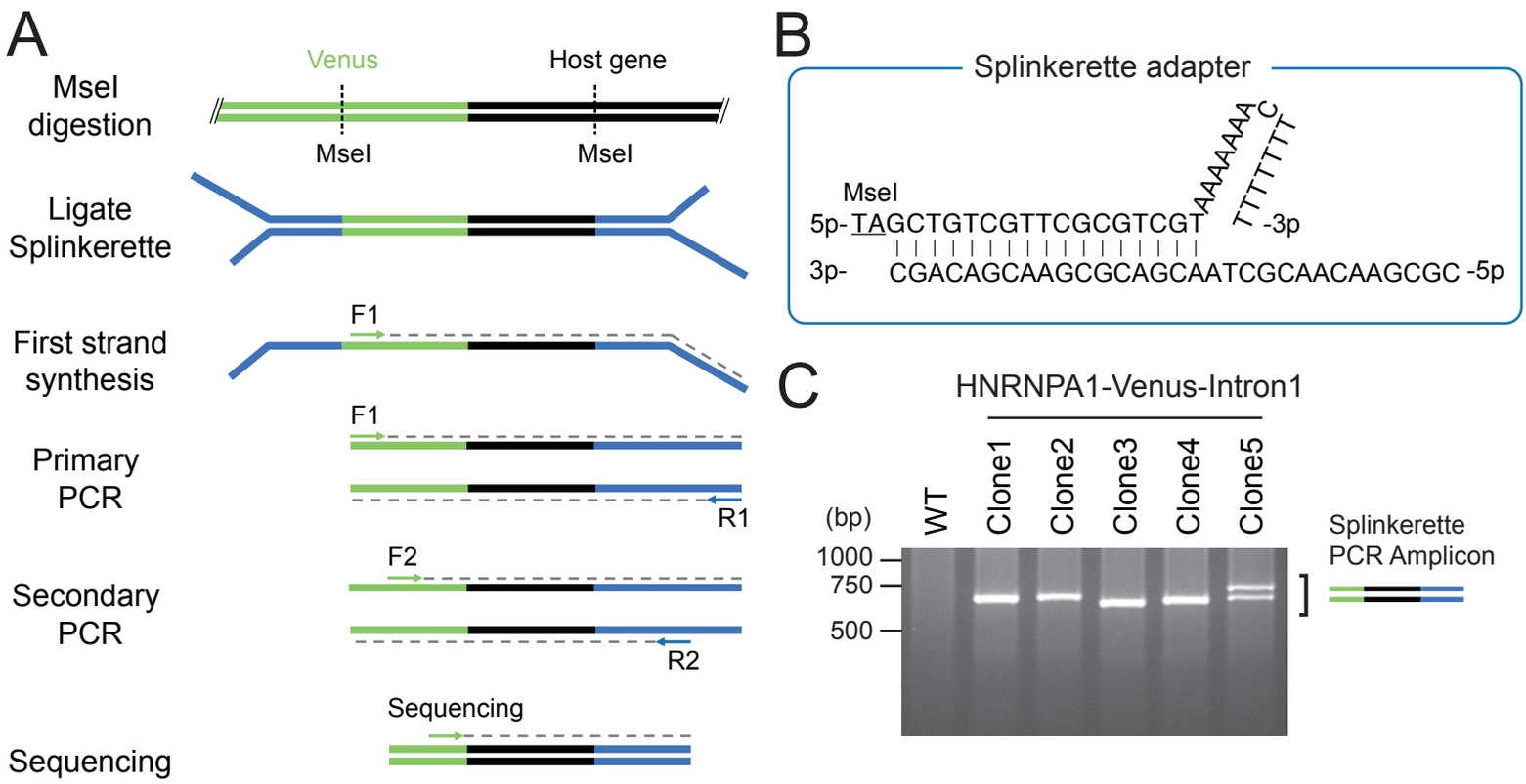




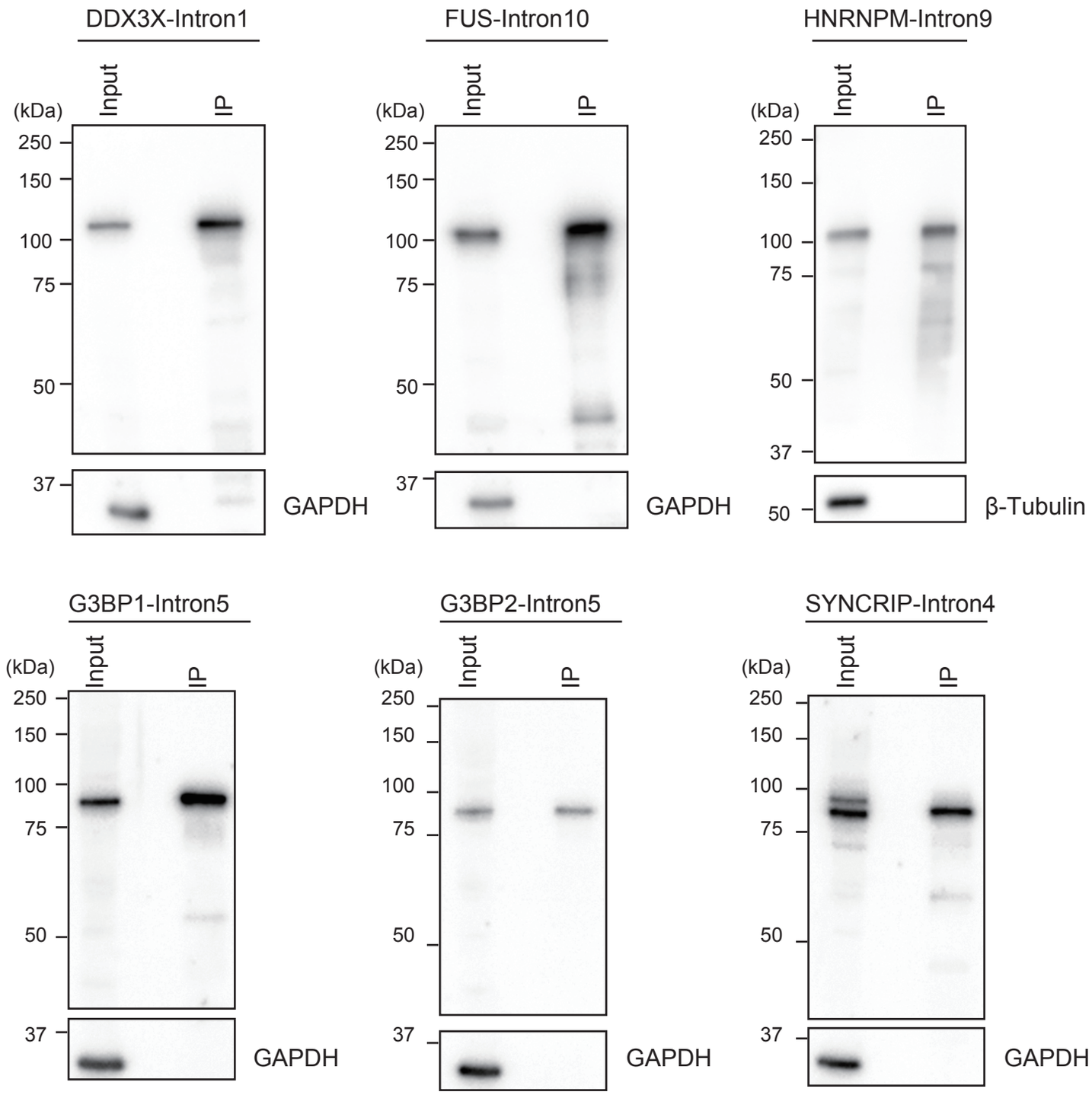
Supplementary Figure 3



Supplementary Figure 4



Supplementary Figure 5

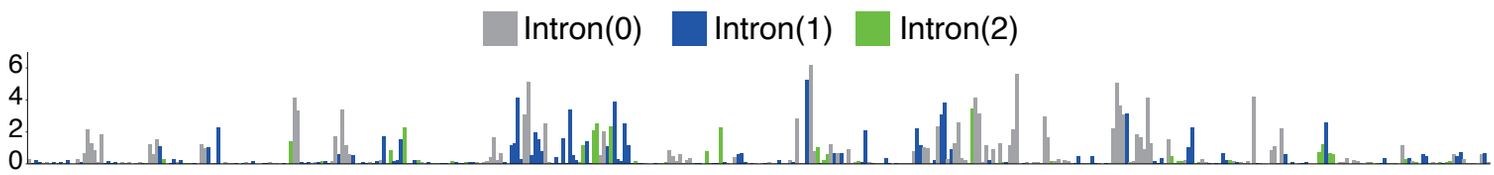


Supplementary Figure 6

A

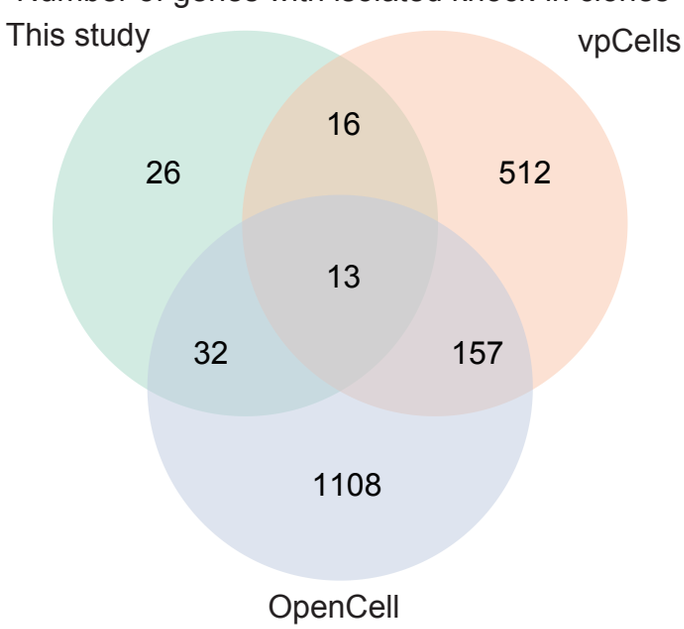
117 genes 425 introns (**Bold**: Venus positivity rate in miRFP670 positive (%) >= 0.1)

ACLY	**DDX3X**	**EIF4A1**	HDLBP	**ILF2**	**NOP56**	**PTBP1**	**SFPQ**	TIA1
ACTN4	DDX5	EIF4B	**HMGA1**	**KHDRBS1**	NOP58	**PTGES3**	SLC25A3	TIAL1
ATP5F1A	DDX17	**EIF4G1**	**HNRNPA1**	**KHSRP**	**NPM1**	**PTMA**	SLC25A5	**TRA2B**
ATP5F1C	**DDX21**	**EIF4H**	**HNRNPA2B1**	**KPNA2**	**P4HB**	**PUF60**	**SNRNP70**	**TRIM28**
ATP5PO	**DEK**	**EIF5**	**HNRNPA3**	LARP1	**PABPC1**	**RACK1**	SNRNP200	**TUBA1C**
AURKAIP1	DHX9	**EIF5A**	**HNRNPAB**	**LDHA**	PABPN1	**RAN**	**SNRPB**	**U2AF2**
AURKB	DKC1	**ENO1**	**HNRNPC**	**LDHB**	**PARP1**	**RBBP4**	**SNRPD1**	**UBA1**
BCLAF1	DNAJC7	**EWSR1**	**HNRNPD**	LUC7L2	**PCBP2**	**RBBP7**	SNRPD2	UBAP2L
BZW1	EBNA1BP2	FARSA	**HNRNPDL**	MATR3	PCNP	**RBM3**	SNRPD3	VCP
C1QBP	**EEF1G**	FASN	**HNRNPH1**	MCM3	**PDAP1**	RBM17	SON	VDAC2
CCT8	**EEF2**	**FBL**	**HNRNPH3**	**MCM4**	PDCD6IP	**RBM39**	SRM	VIM
CDC37	EIF3A	**FUBP1**	**HNRNPK**	MCM5	PDHA1	**RBMX**	SRRT	**XRCC5**
CFL1	EIF3B	**FUS**	**HNRNPL**	MLF2	PES1	**RNPS1**	**SRSF3**	**XRCC6**
CLNS1A	EIF3D	**G3BP1**	**HNRNPM**	MTCH2	PHB2	**RSL1D1**	**SRSF7**	**YBX1**
CSDE1	EIF3E	G3BP2	HNRNPR	**MYL6**	**PKM**	RTRAF	**SRSF9**	YBX3
CSE1L	**EIF3G**	**GAPDH**	**HNRNPU**	**NACA**	**PPIA**	SEPTIN2	SSBP1	**YWHAE**
CTPS1	EIF3H	GTF2I	HNRNPUL1	**NAP1L1**	**PRMT1**	**SF1**	**SSRP1**	ZRANB2
CYC1	EIF3K	H33A	**HSPA8**	**NCL**	PSMC3	**SF3A3**	**SYNCRIP**	
DAZAP1	EIF3L	HDAC1	**HSPD1**	**NOLC1**	**PSMD2**	**SF3B1**	TAF15	
DDX1	EIF3M	**HDAC2**	IARS1	**NONO**	PSMD6	**SF3B2**	**TARDBP**	



B

Number of genes with isolated knock-in clones



Supplementary Figure 7