Nucleic Acids Research, 2024, **52**, e68
https://doi.org/10.1093/nar/gkae553
Advance access publication date: 26 June 2024
Methods

protoSpaceJAM: an open-source, customizable and web-accessible design platform for CRISPR/Cas insertional knock-in

Duo Peng *, Madhuri Vangipuram, Joan Wong and Manuel D. Leonetti *

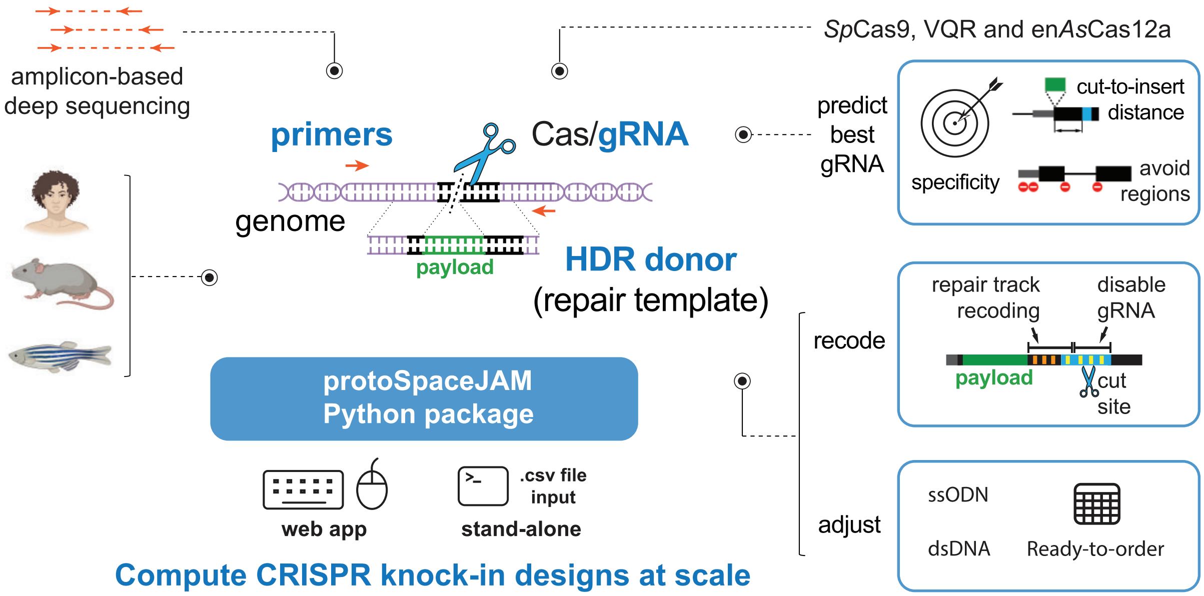
Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

*To whom correspondence should be addressed. Tel: +1 707 5822172; Email: manuel.leonetti@czbiohub.org
Correspondence may also be addressed to Duo Peng. Email: duo.peng@czbiohub.org

Abstract

CRISPR/Cas-mediated knock-in of DNA sequences enables precise genome engineering for research and therapeutic applications. However, designing effective guide RNAs (gRNAs) and homology-directed repair (HDR) donors remains a bottleneck. Here, we present protoSpaceJAM, an open-source algorithm to automate and optimize gRNA and HDR donor design for CRISPR/Cas insertional knock-in experiments, currently supporting SpCas9, SpCas9-VQR and enAsCas12a Cas enzymes. protoSpaceJAM utilizes biological rules to rank gRNAs based on specificity, distance to insertion site, and position relative to regulatory regions. protoSpaceJAM can introduce 'recoding' mutations (silent mutations and mutations in non-coding sequences) in HDR donors to prevent re-cutting and increase knock-in efficiency. Users can customize parameters and design double-stranded or single-stranded donors. We validated protoSpaceJAM's design rules by demonstrating increased knock-in efficiency with recoding mutations and optimal strand selection for single-stranded donors. An additional module enables the design of genotyping primers for deep sequencing of edited alleles. Overall, protoSpaceJAM streamlines and optimizes CRISPR knock-in experimental design in a flexible and modular manner to benefit diverse research and therapeutic applications. protoSpaceJAM is available open-source as an interactive web tool at protospacejam.czbiohub.org or as a standalone Python package at github.com/czbiohub-sf/protoSpaceJAM.

Graphical abstract



Introduction

The development of gene editing technologies, fueled by the discovery of CRISPR/Cas systems, has transformed our ability to manipulate genomes. A key goal of gene editing is to add new features to the genome via the controllable insertion ('knock-in') of functional payloads. Applications of CRISPR knock-in span both research and clinical programs, including the systematic characterization of gene function using fluorescent protein tags (1,2), the introduction of genetic variants for the elucidation of disease mechanisms (3), or the integration of chimeric antigen receptor payloads in T cells (CAR-T) for immunotherapy (4). Reflecting the breadth of these

Received: October 13, 2023. **Revised:** June 6, 2024. **Editorial Decision:** June 11, 2024. **Accepted:** June 17, 2024

applications, many CRISPR/Cas-mediated approaches have been developed to enable knock-in in a variety of contexts (reviewed in (5)). For example, homology-independent methods that rely on non-homologous end-joining (NHEJ) enable knock-in in non-dividing cells (6), while prime editing allows knock-in in the absence of double-strand breaks (7). The most common experimental approach for site-specific knock-in leverages homology-directed repair (HDR) (8). In HDR-based knock-in, a site-specific guide RNA (gRNA) is used to recruit CRISPR/Cas at a genomic target and induce a double-strand break, while an exogenous DNA sequence ('donor') is provided to direct the integration of the desired payload. In the donor, the payload is flanked by sequences homologous to the target site ('homology arms') that can template DNA repair by co-opting endogenous repair pathways (Figure 1A). Because HDR is precise, it remains the most widely used method for genomic knock-in, although its use is restricted to dividing cells (8). Technical advances are rapidly increasing the efficiency of HDR-based knock-in in key cell types such as human stem cells (9–11) or primary T cells (12). In addition, CRISPR/Cas knock-in is now entering the clinic for CAR-T therapy (4).

The design of a CRISPR/Cas HDR-based insertional knock-in experiment (subsequently referred to as 'knock-in') is conceptually simple and involves two key components. For each edit, a guide RNA sequence ('protospacer') targeting the desired genomic region must be chosen, and the sequence of a HDR donor template must be constructed (Figure 1A). In recent years, a set of design rules for both gRNA and donor has been established to optimize knock-in efficiency (13–18). This creates an opportunity to develop an algorithm to programmatically design reagents and streamline knock-in experiments. Because of the large breadth of applications of knock-in across genomic, cell biology, or clinical research, such a design algorithm would benefit a large community. Ideally, this algorithm would be (i) accessible and simple to use, enabling researchers from various fields to harness and utilize knock-in technologies; (ii) customizable, giving the user full control over the parameters used in the design choices and tailored for each application and (iii) fully open-source, with well-documented source code built in a modular format so that parts of the algorithm could be easily reused or modified by others to pave the way for new applications.

Here, we present protoSpaceJAM, an easy-to-use, customizable, open-source and scalable algorithm for CRISPR/Cas insertional knock-in. protoSpaceJAM utilizes a state-of-the-art set of rules for gRNA and HDR donor design and enables the user to fine-tune parameters for each design (including the length of homology arms, the introduction of 'recoding' mutations such as silent mutations and mutations in non-coding sequences, or the avoidance of specific sequence features), while providing sensible default options for novice users. protoSpaceJAM allows users to choose between double-stranded DNA (dsDNA) and single-stranded oligodeoxynucleotide (ssODN) forms of HDR donors, with rules tailored to each donor type to optimize knock-in efficiency and/or facilitate chemical DNA synthesis. A companion algorithm, GenoPrimer, designs sequencing primers to simplify the genotyping of edited alleles by deep sequencing. protoSpaceJAM is currently built for experiments utilizing three commonly used Cas enzymes with different Protospacer Adjacent Motifs (PAMs): *S. pyogenes* SpCas9 (NGG PAM), its SpCas9-VQR variant (NGA PAM) (19) and *Acidaminococcus sp.* AsCa12a (TTTV PAM) (20). protoSpaceJAM supports insertional knock-in applications in which a payload sequence can be inserted into any coordinate in the human, mouse, or zebrafish genomes (as opposed to mutational knock-in applications where a native sequence is replaced by another). Documented source code is available on Github at github.com/czbiohub-sf/protospacejam to enable and facilitate the introduction of other desired features (also archived on Zenodo as of the time of writing: doi.org/10.5281/zenodo.11057858). Finally, protoSpaceJAM is available as a stand-alone Python package as well as a user-friendly web application (protospacejam.czbiohub.org) to catalyze broader accessibility and usage.

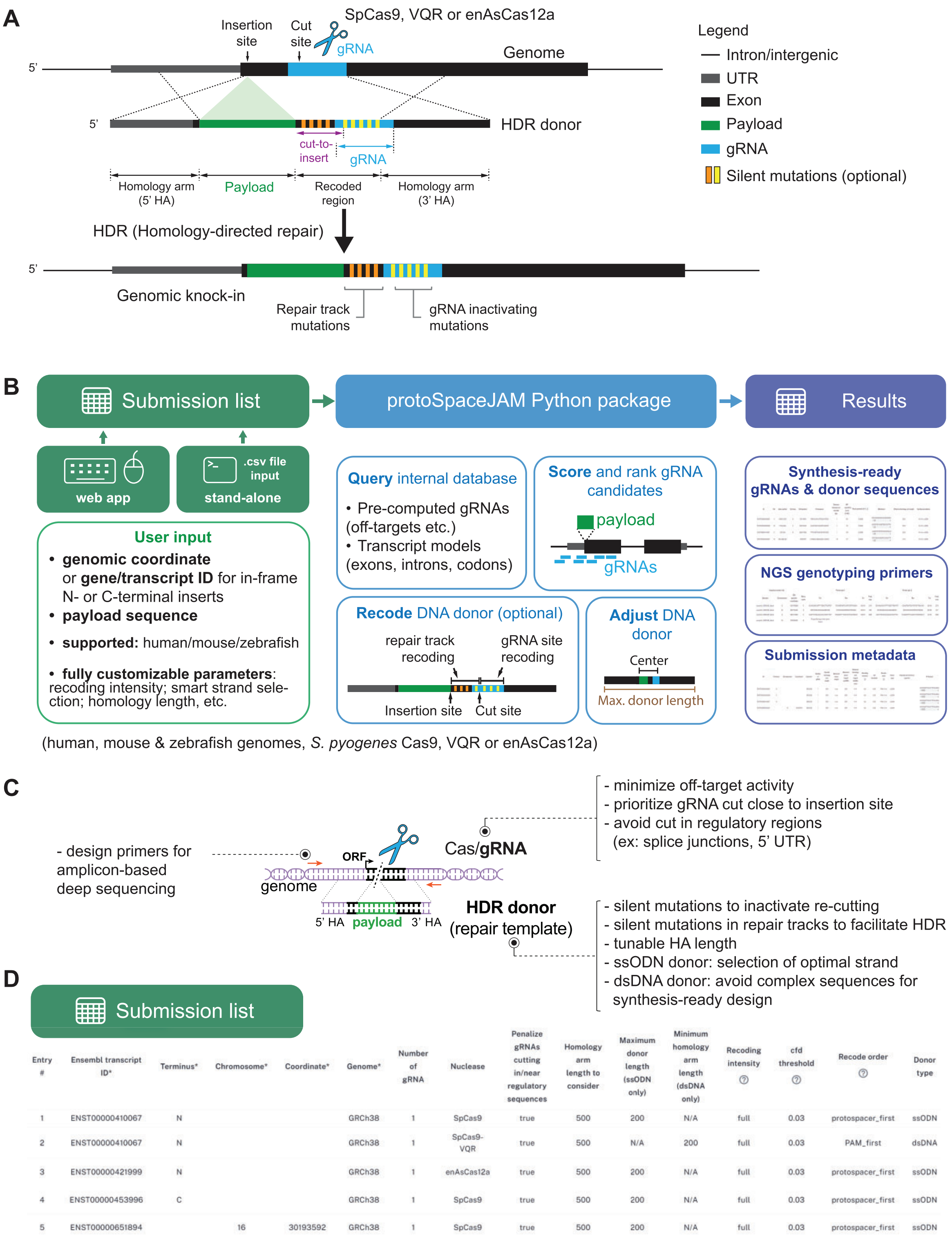
Materials and methods

Genome datasets

The following reference genomes were used: GRCh38 for human, GRCm39 for mouse, and GRCz11 for zebrafish. Genome annotations for human, mouse, and zebrafish were downloaded as GFF3 files from ftp.ensembl.org/pub/release-109. Custom scripts (github.com/czbiohub-sf/protoSpaceJAM/blob/main/protoSpaceJAM/precompute/scripts/extract_gene_models_info.py) were used to parse the GFF3 files to extract and store the boundaries of UTRs, exons, and introns as genomic coordinates for all annotated transcripts. The reading frame of each position in coding sequences was extracted and stored for every annotated transcript. Sequences near splice junctions were extracted using scripts available at github.com/duopeng/JuncSeq and sequence logo plots were generated using weblogo v3.7.12 (21).

gRNA search and off-target score calculation

protoSpaceJAM uses precomputed gRNA information to reduce processing time. Precomputed gRNAs for the human, mouse, and zebrafish genomes are available at github.com/czbiohub-sf/protoSpaceJAM#download-and-unzip-pre-computed-data. The computer code and workflow to precompute gRNA for any given genome is available at: github.com/czbiohub-sf/protoSpaceJAM/tree/main/protoSpaceJAM/precompute. Briefly, all possible protospacers for *S. pyogenes* Cas9 (SpCas9), VQR-SpCas9 and enhanced *Acidaminococcus* sp. Cas12a (enAsCas12a) are enumerated for the entire genome (using PAMs NGG, NGA and TTTV, respectively). Code from the CRISPOR tool (22) was adapted to calculate target specificity score for each protospacer as follows. First, each protospacer sequence is aligned to the genome using the Burrows-Wheeler Aligner (bwa) (23), allowing up to four mismatches. The non-default parameters employed in this alignment process included '-o 0', '-m 2000000', '-n 4', '-k 4', '-N' and '-l 20'. Second, genomic off-target matches are analyzed for the presence of PAMs immediately to the 3′ of the protospacer (considering the following off-target PAMs; SpCas9: NGG, NGA, NAG; VQR-Cas9: NGA, NGG; Cas12a: TTTN). Third, for each off-target match that includes a PAM, the MIT score (13) is computed. The MIT score predicts the propensity of gRNAs binding to off-targets factoring in the mismatch location in the protospacer. Finally, an aggregated MIT score is computed for each protospacer using the following formula: 100 / (100 + sum of off-target MIT scores).


A
Insertion site
Cut site
SpCas9, VQR or enAsCas12a
gRNA
Genome
5'
5'
HDR donor
cut-to-insert
gRNA
Homology arm (5' HA)
Payload
Recoded region
Homology arm (3' HA)
HDR (Homology-directed repair)
5'
Genomic knock-in
Repair track mutations
gRNA inactivating mutations
Legend
— Intron/intergenic
UTR
Exon
Payload
gRNA
Silent mutations (optional)
B
Submission list
protoSpaceJAM Python package
Results
web app
.csv file input
stand-alone
User input
• genomic coordinate or gene/transcript ID for in-frame N- or C-terminal inserts
• payload sequence
• supported: human/mouse/zebrafish
• fully customizable parameters: recoding intensity; smart strand selection; homology length, etc.
Query internal database
• Pre-computed gRNAs (off-targets etc.)
• Transcript models (exons, introns, codons)
Score and rank gRNA candidates
payload
gRNAs
Recode DNA donor (optional)
repair track recoding
gRNA site recoding
Insertion site
Cut site
Adjust DNA donor
Center
Max. donor length
Synthesis-ready gRNAs & donor sequences
NGS genotyping primers
Submission metadata
(human, mouse & zebrafish genomes, S. pyogenes Cas9, VQR or enAsCas12a)
C
- design primers for amplicon-based deep sequencing
ORF
genome
Cas/gRNA
- minimize off-target activity
- prioritize gRNA cut close to insertion site
- avoid cut in regulatory regions (ex: splice junctions, 5' UTR)
5' HA
payload
3' HA
HDR donor (repair template)
- silent mutations to inactivate re-cutting
- silent mutations in repair tracks to facilitate HDR
- tunable HA length
- ssODN donor: selection of optimal strand
- dsDNA donor: avoid complex sequences for synthesis-ready design
D
Submission list

Entry #	Ensembl transcript ID*	Terminus*	Chromosome*	Coordinate*	Genome*	Number of gRNA	Nuclease	Penalize gRNAs cutting in/near regulatory sequences	Homology arm length to consider	Maximum donor length (ssODN only)	Minimum homology arm length (dsDNA only)	Recoding intensity ⓘ	cfd threshold ⓘ	Recode order ⓘ	Donor type
1	ENST00000410067	N			GRCh38	1	SpCas9	true	500	200	N/A	full	0.03	protospacer_first	ssODN
2	ENST00000410067	N			GRCh38	1	SpCas9-VQR	true	500	N/A	200	full	0.03	PAM_first	dsDNA
3	ENST00000421999	N			GRCh38	1	enAsCas12a	true	500	200	N/A	full	0.03	protospacer_first	ssODN
4	ENST00000453996	C			GRCh38	1	SpCas9	true	500	200	N/A	full	0.03	protospacer_first	ssODN
5	ENST00000651894		16	30193592	GRCh38	1	SpCas9	true	500	200	N/A	full	0.03	protospacer_first	ssODN

Figure 1. Key concepts, flowchart, and tunable parameters in protoSpaceJAM. (**A**) Key concepts for CRISPR knock-in design. For each insertion, a gRNA (blue) targeting a genomic region and an HDR donor sequence that templates payload (green) integration must be designed. To increase knock-in efficiency, the HDR donor may contain silent mutations to protect the HDR product from further CRISPR-induced DNA strand breaks ('re-cutting' mutations, yellow stripes) and silent mutations in the 'cut-to-insert' region between the double-strand break and payload integration sites ('repair track' mutations, orange stripes). Recoded regions are not considered part of the effective homology arms. (**B**) Flowchart of protoSpaceJAM components. protoSpaceJAM can be used either as a standalone Python package or via a user-friendly web app. (**C**) Tunable design parameters in protoSpaceJAM. (**D**) Screenshot of the submission list.

gRNA scoring

The strategy for gRNA scoring is described in detail in the main text. All corresponding code can be found at github.com/czbiohub-sf/protoSpaceJAM/blob/main/protoSpaceJAM/util/utils.py.

Primer design with GenoPrimer

Amplification primers for genotype analysis were designed using GenoPrimer, a custom-built open-source pipeline available at github.com/czbiohub-sf/GenoPrimer. Briefly, the pipeline takes a genomic coordinate position as input (typically the CRISPR knock-in edit site), and it then extracts 430 or 4000 base pairs (bp) of genomic sequences centered on the input coordinate for short- or long-read mode, respectively. From the extracted genomic sequence, candidate primers are enumerated and filtered with Primer3 (24) using a set of thermodynamics filters, a positional filter, and a product size filter. The thermodynamics filters are default for Primer3 and include but are not limited to the following: minimum, optimum, and maximum melting temperatures (T_m) of 57, 60 and 63°C, respectively; a maximum difference of 3°C in T_m; GC content between 20% and 80%; and primer lengths of at least 18 bp, with an optimum of 20 bp and a maximum of 25 bp. Primer-dimers are minimized by application of default thresholds for both 3′ self-complementary and 3′ pair-complementary binding. The positional filter was configured using the 'SEQUENCE_EXCLUDED_REGION' directive to remove primers too proximal to the CRISPR knock-in edit site, with the minimum distance set to 100 and 1000 bp for short- and long-read modes, respectively. The PCR product size filter is configured using the 'PRIMER_PRODUCT_SIZE_RANGE' directive, and it is set to 250–350 and 3300–3700 bp for short- and long-read modes, respectively.

Candidate primers that successfully pass all three filters imposed by Primer3 are further analyzed for unintended PCR products in the target genome. All possible annealing sites in the genome are identified using Bowtie (25) by default, with the following custom parameters: '-k 1000' and '-v 3'. Whereas Bowtie limits alignments to a maximum of three mismatches, BLAST is an alternative mapping program that allows more than three mismatches at the cost of computational speed; this option can be selected by passing the argument '–aligner BLAST' to GenoPrimer with the following custom parameters: '-task blastn-short', '-max_hsps 2000', 'perc_identity 75'. Annealing sites are removed if the 3′ end of the primer sequence does not match the genome. The way GenoPrimer checks for off-target primer binding sites differs from the widely used PrimerBlast algorithm (26) in two ways. First, when using the Bowtie aligner, sites containing 4 or more mismatches to a given primer sequence will not be considered in the off-target search, compared to 6 or more mismatches for PrimerBlast. In the web application version of GenoPrimer, we use Bowtie to prioritize computing speed. In the standalone Python code, the BLAST + aligner can alternatively be used as described above, and sites containing five or more mismatches to a given primer sequence will not be considered in the off-target search. Second, GenoPrimer requires a perfect match at the final 3′-end base for an off-target site to be considered, while PrimerBlast requires a match in at least four of the five final 3′-end bases.

Unintended PCR products are identified when a primer pair possesses annealing sites on opposite DNA strands with their 3′ ends facing each other, with a maximum predicted amplicon size set to 6 kb by default. GenoPrimer checks whether unintended PCR products can be formed between forward + reverse, forward + forward, and reverse + reverse primers of the same pair; if so, the primer pair is removed from the list of candidates. If no successful primer pairs can be found, a series of six attempts will be made to find the next-best primer pairs. In the first, second, and third attempts, the maximum difference in Tm between paired primers is increased by 1°C. In the second, fourth, and sixth attempts, the upper limit of the PCR product size is increased by 80 bp and 300 bp for short- and long-read mode, respectively. To accommodate the increased PCR product size range, the extracted genomic sequences are extended on both ends by 40 and 150 bp for short- and long-read mode, respectively.

protoSpaceJAM Python package and web portal

protoSpaceJAM is available as a pip-installable Python package available at github.com/czbiohub-sf/protoSpaceJAM. Users have the flexibility to fine-tune the underlying algorithm and customize parameters. Specific documentation to do so is available at github.com/czbiohub-sf/protoSpaceJAM/wiki. For example, one can adjust the mathematical formulas responsible for calculating gRNA scoring weights and redefine how synonymous codons are chosen for silent mutations.

Local versions of protoSpaceJAM require precomputed gRNA information, which can be automatically downloaded and set up for the human, mouse, and zebrafish genomes following the installation instructions at github.com/czbiohub-sf/protoSpaceJAM/blob/main/README.md. Users can precompute gRNA information for genomes of their choice following a set of instructions and scripts available at github.com/czbiohub-sf/protoSpaceJAM/tree/main/protoSpaceJAM/precompute.

An interactive web tool is available at protospacejam.czbiohub.org. All source code for the web tool is available at github.com/czbiohub-sf/protoSpaceJAM-portal, which also includes instructions to set up local versions of the interactive web tool for specific applications.

Cell culture

HEK293T

Human HEK-293T cells (ATCC #CRL-3216) were cultured in Gibco DMEM, High Glucose, GlutaMAX Supplement media (Thermo Fisher scientific #10566024) with 10% fetal bovine serum (Omega Scientific #FB-11). The cells were maintained at 37°C and 5% CO_2 and were passaged upon reaching ∼80% confluency using 0.05% Trypsin–EDTA (Thermo Fisher Scientific #25300120).

Human induced pluripotent stem cells (hiPSCs)

We used a variant of the well-characterized WTC11 hiPSC line (Coriell #GM25256) constitutively expressing the $GFP_{1\text{-}10}$ split-GFP construct (27). $WTC_{GFP\ 1\text{–}10}$ iPSCs were passaged on plates pre-coated with vitronectin (Thermo Fisher Scientific #A14700) using MteSR Plus media (Stem Cell Technologies # 100-0276) and RevitaCell supplement (Thermo Fisher Scientific #A26445-01). The cells were maintained at 37°C and 5% CO_2 and were passaged upon reaching ∼80% confluency using Accutase (Stem Cell Technologies #7920).

Genome editing in HEK293T cells with fluorescent payloads and flow cytometry (cf. Figures 2C, 3C, 4D)

For monitoring of HDR efficiency, we inserted sequences encoding fluorescent proteins in HEK-293T cells using methods described in (1). Briefly, ribonucleoprotein (RNP) *S. pyogenes* Cas9/gRNA complexes were prepared *in vitro*, mixed with HDR donor templates, and electroporated into HEK-293T cells by nucleofection using SF solution (Lonza #V4SC-2096) and the CM-130 program. Five days post-nucleofection, the distribution of fluorescence signal in each target cell population was measured by analytical flow cytometry on a FACSymphony instrument (BD Biosciences). Flow cytometry data analysis was performed using the FlowJo software (BD Biosciences). All gRNA and HDR donor sequences used in this study are available in Supplementary Table S1, together with the corresponding numbers of % fluorescent cells.

Genome editing in hiPSC cells with fluorescent payloads and flow cytometry (cf. Supplementary Figure S3)

For monitoring of HDR efficiency, we inserted sequences encoding the GFP11 fluorescent tag (27) in $\mathrm{WTC}_{\mathrm{GFP}\ 1\text{–}10}$ hIPSC cells using methods described in (11). Briefly, ribonucleoprotein (RNP) *S. pyogenes* Cas9/gRNA complexes were prepared *in vitro*, mixed with HDR donor templates, and electroporated into $\mathrm{WTC}_{\mathrm{GFP}\ 1\text{–}10}$ cells by nucleofection using P3 solution (Lonza #V4SP-3096) and the CA-137 program. Nucleofected cells were plated in StemFlex media (Thermo Fisher Scientific #A3349401) supplemented with 1× CloneR-2 recovery reagent (Stem Cell Technologies, #100-0691) and 1 μM HDR enhancer v2 (Integrated DNA Technologies #10007910), and they were maintained at 32°C and 5% CO_2 for 2 days (cold shock). Cells were grown at 37°C and 5% CO_2 for an additional 3 days, at which point the distribution of fluorescence signal in each cell population was measured by analytical flow cytometry on a FACSymphony instrument (BD Biosciences). Flow cytometry data analysis was performed using the FlowJo software (BD Biosciences). All gRNA and HDR donor sequences used in this study are available in Supplementary Table S1, together with the corresponding numbers of % fluorescent cells.

Genotype analysis of OpenCell lines (cf. Figure 2C)

The genotype of 271 cell lines from the OpenCell project was analyzed by next-generation sequencing as described in (1). In these experiments, a split-mNeongreen fluorescent payload was inserted in 271 different protein-coding genes in HEK-293T cells by nucleofection of Cas9/gRNA RNPs and single-stranded DNA donors. For each targeted gene, the genotype of a polyclonal pool of ∼20 000 cells was characterized 5 days post-nucleofection in the absence of any selection, so that HDR rates could be directly measured. For each cell pool, genomic DNA was first extracted by cell lysis using QuickExtract DNA Extraction Solution (Lucigen), from which target gene-specific amplicon libraries were prepared using a two-step PCR protocol as described in (1): the first PCR amplifies the target genomic locus and adds universal amplification handle sequences, while the second PCR introduces indexed Illumina barcodes using these universal handles. Barcoded amplicons were analyzed using capillary electrophoresis (Fragment Analyzer, Agilent #DNF-474-0500), pooled, and purified using solid-phase reversible immobilization magnetic beads. Sequencing was performed on an Illumina Miseq V3 platform (2 × 300 bp paired-end reads) using standard P5/P7 primers. Genotype analysis was performed using CRISPRESSO2 (28) to quantify HDR rate for each target gene (defined as the percentage of HDR alleles out of all non-wild type alleles sequenced, to normalize for the cutting efficiency of each gRNA). gRNA, HDR donor, and primer sequences and genotype analysis for all targets are found in Supplementary Table S2.

GenoPrimer amplification test (cf. Figure 5B)

To test GenoPrimer's ability to design amplification primers for genotype analysis, we designed primer pairs for 94 separate genes in short-read mode. Products were amplified from wild-type genomic DNA purified from wild-type HEK-293T cells (New England Biolabs #T3010S) using a touch-down PCR strategy. 40-μl PCR reactions were set using 2× KAPA HiFi Hotstart reagents (Roche #KK2602) with 10 ng genomic DNA, 2 μM of each primer, and betaine to 1 M final concentration. PCR conditions: 95°C for 3min; 2 cycles of {98°C for 20 s, 72°C for 15 s, 72°C for 20s}, 2 cycles of {98°C for 20 s, 71°C for 15 s, 72°C for 20 s}, 2 cycles of {98°C for 20 s, 70°C for 15 s, 72°C for 20 s}; 2 cycles of {98°C for 20 s, 69°C for 15 s, 72°C for 20 s}; 2 cycles of {98°C for 20 s, 68°C for 15 s, 72°C for 20 s}; 2 cycles of {98°C for 20 s, 67°C for 15 s, 72°C for 20 s}; 2 cycles of {98°C for 20 s, 66°C for 15 s, 72°C for 20 s}; 2 cycles of {98°C for 20 s, 65°C for 15 s, 72°C for 20 s}; 2 cycles of {98°C for 20 s, 64°C for 15 s, 72°C for 20 s}; 2 cycles of {98°C for 20 s, 63°C for 15 s, 72°C for 20 s}; 10 cycles of {98°C for 20 s, 62°C for 15 s, 72°C for 20 s}; then 72°C for 1 min (final extension); 4°C final. After PCR, the size distribution of amplicons for each target gene was characterized by quantitative capillary electrophoresis (Fragment Analyzer, Agilent #DNF-474-0500). All primer sequences for the GenoPrimer test are found in Supplementary Table S3.

Results

Anatomy of knock-in design and overview of protoSpaceJAM

protoSpaceJAM is a design tool currently developed for insertional HDR-based knock-in applications. In these applications, the genomic insertion of a functional payload is templated by a HDR donor containing the payload flanked by sequences called 'homology arms' that are homologous to the desired insertion site (Figure 1A). For each insertion, two separate components must be designed: a gRNA targeting the genomic region of interest, and the HDR donor sequence itself. In cases where the original protospacer might be preserved within one of the homology arms, introduction of silent 'recoding' mutations may be desirable to inactivate gRNA binding and re-cutting of the knock-in allele (29) and/or to recode the genomic portion located between the desired insertion site and the Cas/gRNA cut site (30) (see Figure 1A; the rationale for 'recoding' mutations is further explained in the next sections).

protoSpaceJAM employs a simple input: the user specifies a genomic insertion site and a payload sequence (Figure 1B). Currently, designs for human, mouse, and zebrafish genomes are supported, together with three commonly used Cas enzymes: *S. pyogenes* SpCas9 (NGG PAM), its SpCas9-

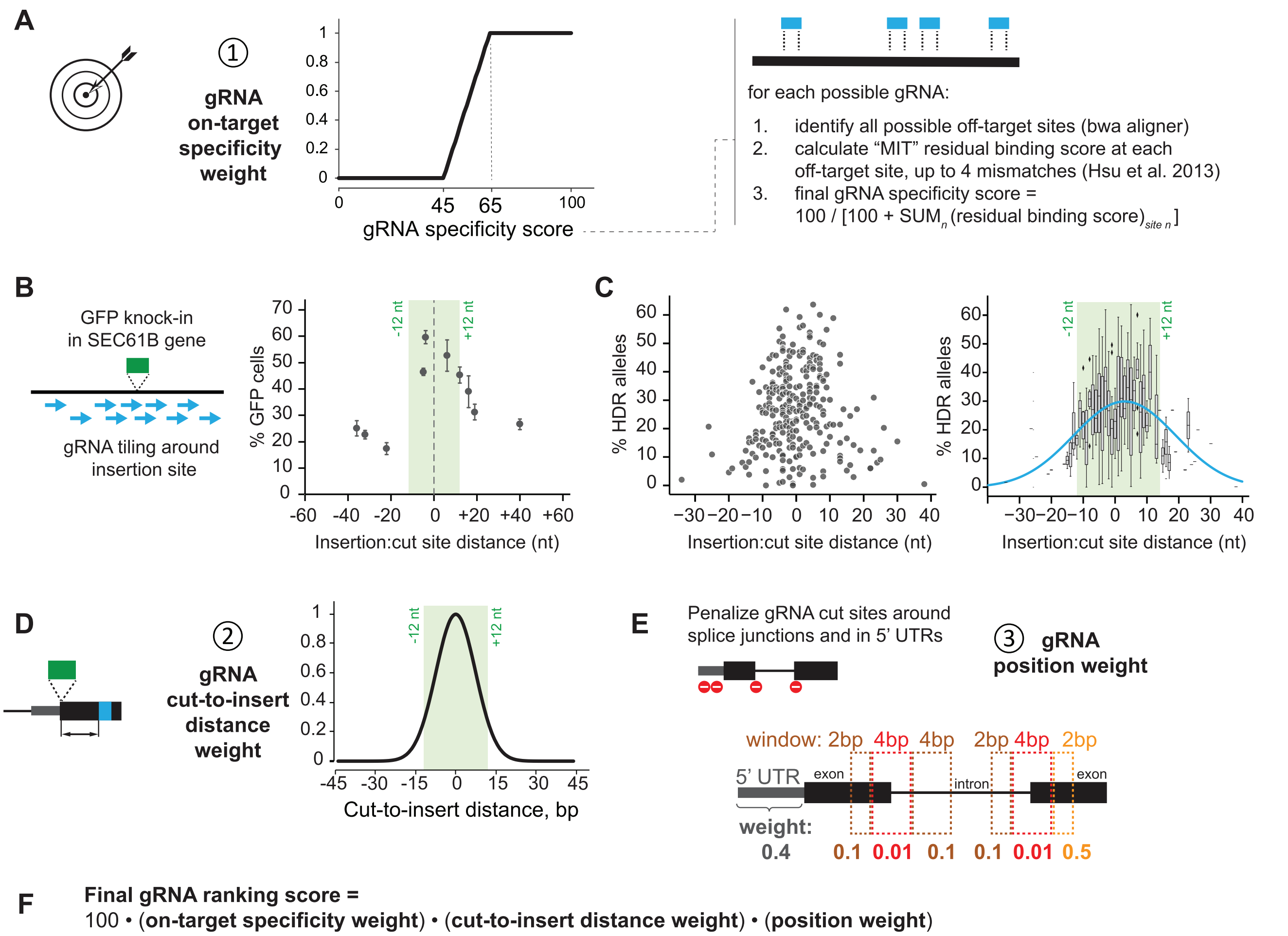
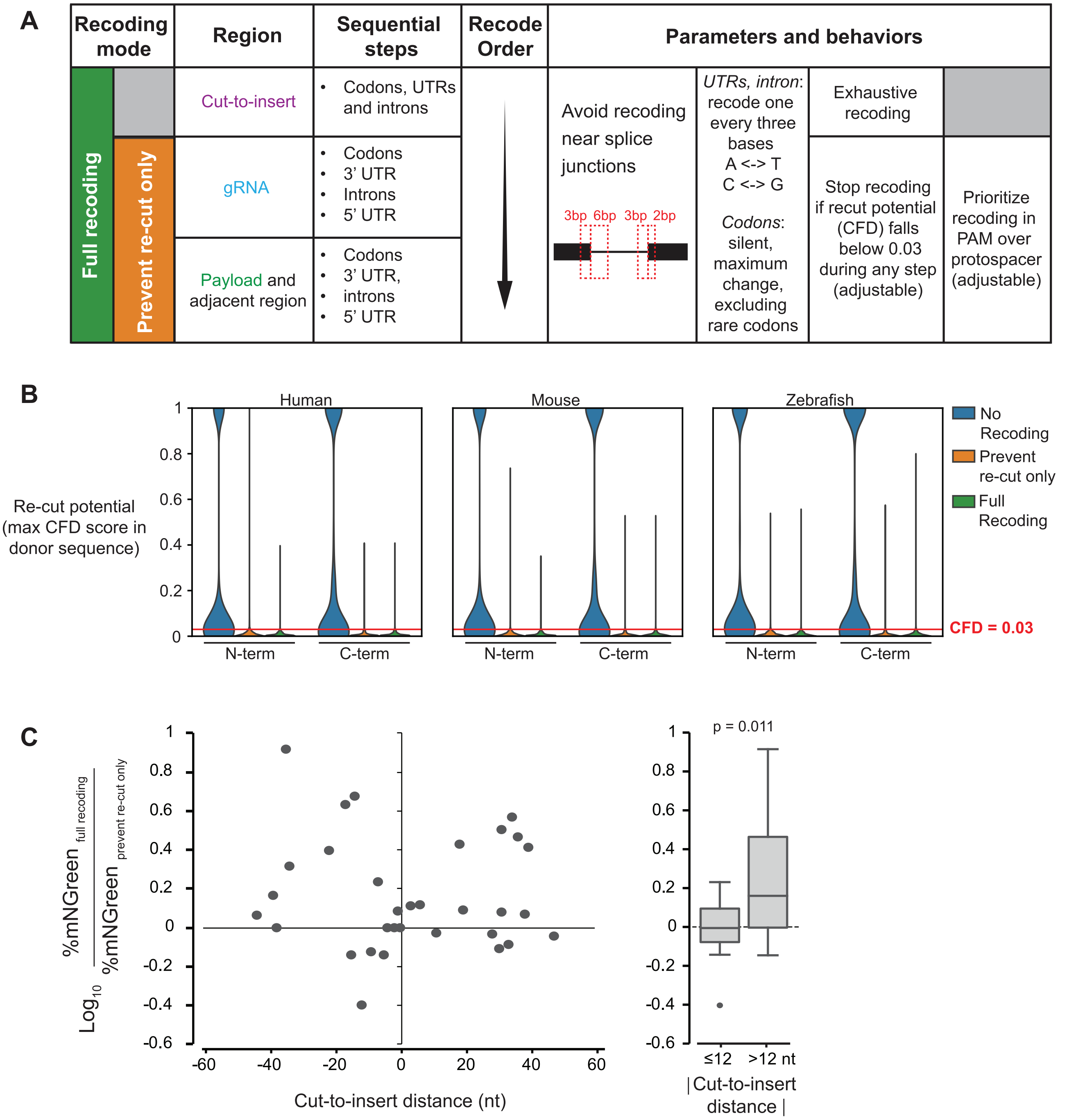


Figure 2. Guide RNA scoring and ranking strategy. (**A**) gRNA specificity is captured by using an 'on-target specificity' weight ranging from 0 to 1, where 0 is for specificity scores below 45 and 1 is for specificity scores above 65, and a linear interpolation is made for scores between 45 and 65. Specificity scores are calculated using a variation of the MIT specificity formula described in CRISPOR (22) that considers any possible off-target gRNA binding sites containing up to 4 mismatches to the original protospacer. (**B**) Payload integration efficiency as measured by the integration of a split-GFP payload (2xGFP11) in the SEC61B locus in HEK293T cells using ssDNA donor templates. The percentage of GFP-positive cells is measured by flow cytometry and serves as a read-out of knock-in efficiency. Graph shows average values (filled circles) from $n = 4$ replicates (error bars show standard deviation). (**C**) HDR efficiency as measured by sequence analysis of targeted alleles. The insertion of a split-mNeonGreen payload was characterized across 271 gene loci in HEK293T cells. The rate of HDR for each gene was characterized by deep sequencing of the targeted alleles. The data are presented as a scatter plot (left) or a box plot (right; boxes represent 25th, 50th, and 75th percentiles, and whiskers represent 1.5 times the interquartile range; outliers are shown as diamonds). An unconstrained Gaussian fit to this distribution is presented as a blue line ($y = 4.3 + 26.6 \times \exp[-(x - 2.9)2/360)$. (**D**) protoSpaceJAM seeks to prioritize gRNAs proximal to the insertion site using a simple 'cut-to-insert distance' Gaussian weight between 0 and 1, with a width of around 12 bp (weight = exp[-distance^2/110]). (**E**) Optional 'gRNA position' weight between 0 and 1 used by protoSpaceJAM to penalize gRNAs targeting 5′ UTRs or splice junctions. (**F**) protoSpaceJAM ranks all possible gRNAs for each knock-in design using the product of all three 'on-target specificity', 'cut-to-insert distance' and 'position' weights in a final composite score.

VQR variant (NAG PAM) (19), and the enhanced version of *Acidaminococcus sp.* AsCa12a (TTTV PAM) (20). Design speed is maximized by building an internal database of pre-computed off-target information of all possible protospacers for all three enzymes in each genome.

Because one of the main applications of knock-in is to insert functional tags in-frame of a protein-coding gene (e.g. for fluorescent protein tagging), the user can also specify an Ensembl transcript ID for a gene of interest and an N- or C-terminal insertion point (for example: ENST00000421999,N). Explicit genomic coordinates can be used for more flexible designs (for example: ENST00000651894,16:30193592 or simply 16:30193592, in the format chromosome_number:position). When a transcript ID is specified, protoSpaceJAM uses the transcript version annotated in Ensembl as of March 2024 (e.g. ENST00000421999.8 for the ENST00000421999 transcript).

The goal of protoSpaceJAM is to streamline the design of both gRNA and donor sequences using a biologically-informed set of rules that are fully described in subsequent sections (summarized in Figure 1C). Importantly, all parameters are fully controlled by the user to maximize utilization across many different applications. Multiple knock-in designs can be handled in parallel by building a submission list, which contains insertion position and payload information as well as the parameter details to be used for each design (Figure 1D). All documented source code is openly available at github.com/czbiohub-sf/protoSpaceJAM, and the tool can be used either as a pip-installable standalone Python package or via a user-friendly web interface at protospacejam.czbiohub.org.



Recoding mode		Region	Sequential steps	Recode Order	Parameters and behaviors			
Full recoding		Cut-to-insert	Codons, UTRs and introns	↓	Avoid recoding near splice junctions	*UTRs, intron*: recode one every three bases A <-> T C <-> G *Codons*: silent, maximum change, excluding rare codons	Exhaustive recoding	
	Prevent re-cut only	gRNA	Codons 3' UTR Introns 5' UTR				Stop recoding if recut potential (CFD) falls below 0.03 during any step (adjustable)	Prioritize recoding in PAM over protospacer (adjustable)
		Payload and adjacent region	Codons 3' UTR, introns 5' UTR					

Figure 3. Mutational recoding in the HDR donor. (**A**) Recoding options in protoSpaceJAM. Two recoding levels are available: 'prevent re-cut only' (orange) introduces silent mutations around the Cas/gRNA binding site to prevent re-cutting of the knock-in allele, while 'full recoding' (green) additionally introduces silent mutations in the cut-to-insert region to prevent DNA repair from resolving before incorporation of the payload sequence. (**B**) Recoding brings the re-cut potential below the desired threshold (CFD < 0.03) in 97.6%, 97.8% and 97.2% of knock-in alleles for insertion of a split-mNeonGreen payload (1) at the N- or C-terminus of all canonical protein-coding transcripts in the human, mouse, or zebrafish genomes (n = 19 643, 24 714 and 25 406, respectively). (**C**) 'Repair track' recoding in the cut-to-insert region affects knock-in efficiency of a split-mNeonGreen payload across 32 human protein-coding genes for which top-scoring gRNAs were located at various cut-to-insert distances. The percentage of GFP-positive cells is measured by flow cytometry and serves as a read-out of knock-in efficiency. P-value: two-sided unpaired Student's t-test.

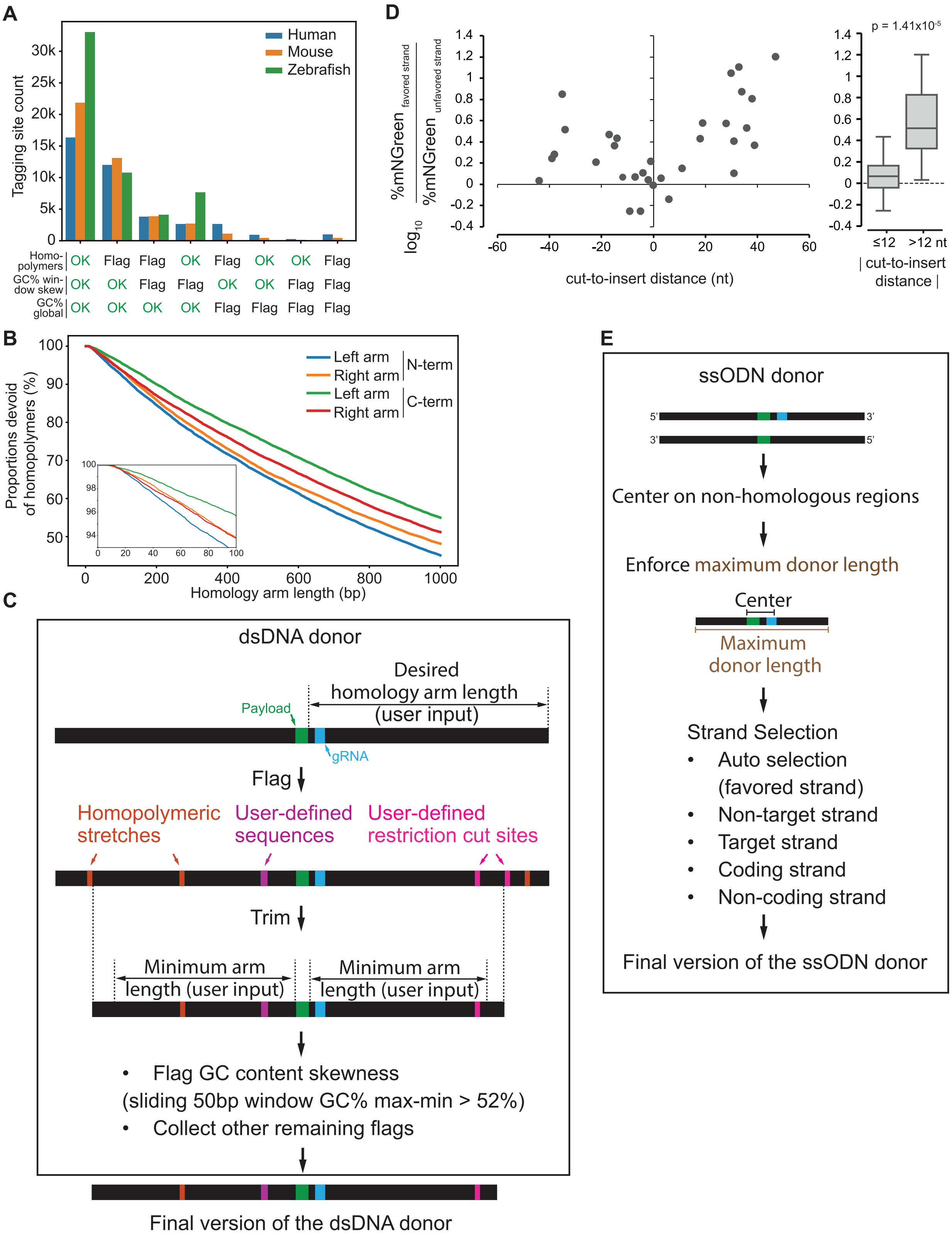
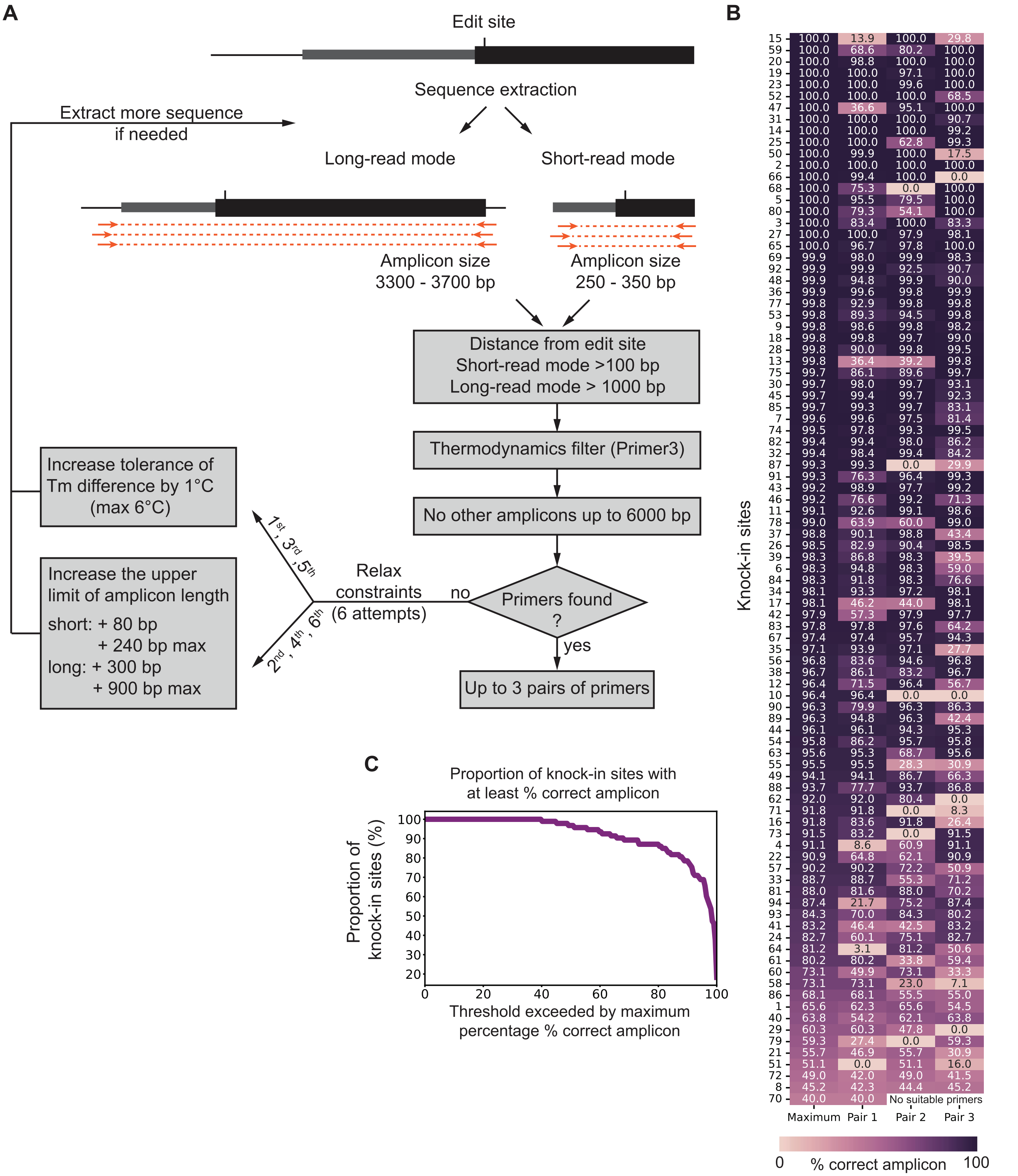


Figure 4. Design parameters for dsDNA and ssODN donors. (**A**) Analysis of 500-bp homology arms flanking the N- and C-terminus tagging sites in all canonical transcripts in human, mouse and zebrafish genomes indicates that homopolymeric runs are the leading problem that would prevent successful synthesis of the dsDNA donor. Homopolymers are defined by homopolymeric runs of 10 or more As and Ts or 6 or more Gs and Cs; GC% window skew is defined by GC content skewness between any sliding window of 50 bp exceeding 52%; GC% global is flagged when the global GC content is below 25% or above 65%. (**B**) Proportions of homology arms devoid of homopolymers as a function of length. N- and C-terminus insertion sites ($n = 39\,286$) in all canonical transcripts in the human genome were analyzed. (**C**) Workflow to remove or flag hard-to-synthesize motifs in dsDNA donor. (**D**) Integration efficiencies of a split-mNeonGreen fluorescent payload across 32 human protein-coding genes for which top-scoring gRNAs were located at various cut-to-insert distances, comparing both strand orientations for each ssODN donor. The percentage of mNeonGreen-positive cells is measured by flow cytometry and serves as a read-out of knock-in efficiency. *P*-value: two-sided unpaired Student's *t*-test. (**E**) Workflow to design synthesis-ready ssODN donor sequences.



Knock-in sites	Maximum	Pair 1	Pair 2	Pair 3
15	100.0	13.9	100.0	29.8
59	100.0	68.6	80.2	100.0
20	100.0	98.8	100.0	100.0
19	100.0	100.0	97.1	100.0
23	100.0	100.0	99.6	100.0
52	100.0	100.0	100.0	68.5
47	100.0	36.6	95.1	100.0
31	100.0	100.0	100.0	90.7
14	100.0	100.0	100.0	99.2
25	100.0	100.0	62.8	99.3
50	100.0	99.9	100.0	17.5
2	100.0	100.0	100.0	100.0
66	100.0	99.4	100.0	0.0
68	100.0	75.3	0.0	100.0
5	100.0	95.5	79.5	100.0
80	100.0	79.3	54.1	100.0
3	100.0	83.4	100.0	83.3
27	100.0	100.0	97.9	98.1
65	100.0	96.7	97.8	100.0
69	99.9	98.0	99.9	98.3
92	99.9	99.9	92.5	90.7
48	99.9	94.8	99.9	90.0
36	99.9	99.6	99.8	99.9
77	99.8	92.9	99.8	99.8
53	99.8	89.3	94.5	99.8
9	99.8	98.6	99.8	98.2
18	99.8	99.8	99.7	99.0
28	99.8	90.0	99.8	99.5
13	99.8	36.4	39.2	99.8
75	99.7	86.1	89.6	99.7
30	99.7	98.0	99.7	93.1
45	99.7	99.4	99.7	92.3
85	99.7	99.3	99.7	83.1
7	99.6	99.6	97.5	81.4
74	99.5	97.8	99.3	99.5
82	99.4	99.4	98.0	86.2
32	99.4	98.4	99.4	84.2
87	99.3	99.3	0.0	29.9
91	99.3	76.3	96.4	99.3
43	99.2	98.9	97.7	99.2
46	99.2	76.6	99.2	71.3
11	99.1	92.6	99.1	98.6
78	99.0	63.9	60.0	99.0
37	98.8	90.1	98.8	43.4
26	98.5	82.9	90.4	98.5
39	98.3	86.8	98.3	39.5
6	98.3	94.8	98.3	59.0
84	98.3	91.8	98.3	76.6
34	98.1	93.3	97.2	98.1
17	98.1	46.2	44.0	98.1
42	97.9	57.3	97.9	97.7
83	97.8	97.8	97.6	64.2
67	97.4	97.4	95.7	94.3
35	97.1	93.9	97.1	27.7
56	96.8	83.6	94.6	96.8
38	96.7	86.1	83.2	96.7
12	96.4	71.5	96.4	56.7
10	96.4	96.4	0.0	0.0
90	96.3	79.9	96.3	86.3
89	96.3	94.8	96.3	42.4
44	96.1	96.1	94.3	95.3
54	95.8	86.2	95.7	95.8
63	95.6	95.3	68.7	95.6
55	95.5	95.5	28.3	30.9
49	94.1	94.1	86.7	66.3
88	93.7	77.7	93.7	86.8
62	92.0	92.0	80.4	0.0
71	91.8	91.8	0.0	8.3
16	91.8	83.6	91.8	26.4
73	91.5	83.2	0.0	91.5
4	91.1	8.6	60.9	91.1
22	90.9	64.8	62.1	90.9
57	90.2	90.2	72.2	50.9
33	88.7	88.7	55.3	71.2
81	88.0	81.6	88.0	70.2
94	87.4	21.7	75.2	87.4
93	84.3	70.0	84.3	80.2
41	83.2	46.4	42.5	83.2
24	82.7	60.1	75.1	82.7
64	81.2	3.1	81.2	50.6
61	80.2	80.2	33.8	59.4
60	73.1	49.9	73.1	33.3
58	73.1	73.1	23.0	7.1
86	68.1	68.1	55.5	55.0
1	65.6	62.3	65.6	54.5
40	63.8	54.2	62.1	63.8
29	60.3	60.3	47.8	0.0
79	59.3	27.4	0.0	59.3
21	55.7	46.9	55.7	30.9
51	51.1	0.0	51.1	16.0
72	49.0	42.0	49.0	41.5
8	45.2	42.3	44.4	45.2
70	40.0	40.0	No suitable primers	

Figure 5. Computational pipeline for design of genotyping primers and experimental validation. (**A**) GenoPrimer predicts primers suitable for amplifying PCR products for short- and long-read sequencing, with default amplicon size ranges of 250–350 bp and 3300–3700 bp, respectively. Primer candidates undergo additional screening using thermodynamic filters implemented in Primer3 (24). To check for unintended PCR products, all potential annealing sites in the genome are searched using Bowtie (or optionally BLAST+). Unintended PCR products can be identified when a primer pair has annealing sites on opposite DNA strands with their 3′ ends oriented towards each other. If no candidate primers remain, GenoPrimer employs a strategy of six sequential attempts to find the next-best primer pairs. This is done by gradually increasing the tolerance for differences in melting temperature (T_m) and extending the upper limit of amplicon length. (**B**) Experimental validation. GenoPrimer was used to design short-read amplification primers for 94 separate loci. At least one primer pair matching the criteria outlined in **(A)** was identified for 93 out of 94 loci (98.9% success rate). Amplification with each primer pair was validated using HEK293T genomic DNA and quantified by capillary electrophoresis. The percentage of amplicons of the correct size over all amplified products is shown for each of the top 3 primer pairs, along with the maximum percentage among all primer pairs for each locus. (**C**) Proportion of knock-in sites for which the percentage of amplicons of the correct size (y-axis) is above a specific threshold (x-axis).

gRNA selection rules

The first step in knock-in design is the selection of a desirable gRNA for optimal efficiency and specificity. To rank all candidate gRNAs for optimal selection, protoSpaceJAM uses a composite ranking score that weighs the on-target specificity of each candidate, the distance between cut and insertion sites, and the position of the gRNA with respect to important gene expression regulatory sequences, namely 5′ untranslated regions (UTRs) and splice sites (Figure 2A–E).

Specificity weight

A key feature of each gRNA is its specificity for a single genomic site, as Cas nucleases might cut at off-target genomic sequences that share significant homology to a given protospacer (31,32). Multiple scoring algorithms have been developed to predict gRNA specificity (reviewed in (33)), including the widely used scoring strategy originally developed by Hsu et al. and referred to as the MIT score (13). protoSpaceJAM uses a variation of the MIT specificity score described in CRISPOR (22) that considers any possible off-target gRNA binding sites containing up to four mismatches to the original protospacer. A residual binding score is assigned to each potential off-target site to evaluate the probability of unintended edits, considering both the position and the density of mismatches. The overall specificity score for a given gRNA is then computed by dividing 100 by the sum of all individual residual binding scores (Figure 2A, right panel). Off-target modification frequencies drop significantly for gRNAs with specificity scores above 45, and they become minimal for scores above 65 (33). For ranking purposes, protoSpaceJAM uses a simple rule to capture gRNA specificity using an 'on-target specificity weight' between 0 and 1, where 0 is assigned to gRNAs with a specificity score below 45 (to severely penalize non-specific guides), 1 is for specificity scores above 65, and a linear interpolation is made for scores between 45 and 65 (Figure 2A).

Cut-to-insert distance weight

For knock-in applications, many studies have shown that the distance between the Cas/gRNA cut site and the desired insertion point is a key parameter that governs integration efficiency (10,15,29,34,35). The further the genomic cut site is from the insertion site, the greater the probability that DNA repair might resolve without payload insertion (because homology to the genome that exists in the donor region between cut and insert will allow DNA repair to resolve prematurely, Supplementary Figure S1A). To test the relationship between integration efficiency and cut-to-insert distance, we measured the integration of a fluorescent payload in the human SEC61B protein-coding gene while tiling gRNAs across the insertion point. Integration efficiency (measured by the percentage of fluorescent cells detected) dropped sharply as a function of cut-to-insert distance, with distances over 12 bp exhibiting the greatest decrease (Figure 2B). Similar relationships have been reported in the literature (36). To further characterize the relationship between HDR efficiency and cut-to-insert distance, we analyzed data from the OpenCell project (1), in which a short fluorescent payload was inserted in a large number of human protein-coding genes, fortuitously spanning a range of cut-to-insert distances. For 271 randomly chosen insertions, the targeted allele was sequenced from edited cell pools and the rate of HDR was quantified (see Materials and methods). This analysis also shows a reduction of HDR efficiency as cut-to-insert distance increases (Figure 2C). As a consequence of these results, protoSpaceJAM uses a scoring weight to prioritize gRNAs with cut sites proximal to the insertion site. By default, protoSpaceJAM uses a 'cut-to-insert distance' Gaussian weight between 0 and 1, with a width centered around 12 bp (Figure 2D).

Additional positional weight

Targeting Cas9/gRNA cuts within important regulatory sequences such as 5′ UTRs or splice junctions should be avoided because the subsequent need to introduce 'recoding' mutations around the cut site (see below) might impact these regulatory regions and alter the endogenous expression of the target gene. 5′ UTRs contain ribosome-binding sequences and play a key role in controlling gene expression (37), while splice junctions contain residues that are universally conserved and cannot be modified (Supplementary Figure S2). Therefore, we defined a 'gRNA position weight' between 0 and 1 to penalize gRNAs targeting these regions (see Figure 2E). Because this optimization strategy is only relevant to some applications, using the positional weight is optional.

Final weight calculation and gRNA scoring

For each knock-in design, protoSpaceJAM ranks all possible gRNAs using the product of the 'on-target specificity', 'cut-to-insert distance', and 'position' weights in a final composite score (Figure 2F). The user can specify the desired number of gRNAs to be returned for each design, prioritized according to their composite score. The detailed weight information for each guide is specified in protoSpaceJAM's table output. Overall, the goal of protoSpaceJAM is to nominate gRNAs that are specific and likely to achieve high knock-in rates while minimally affecting the genetic regulatory patterns of endogenous genes. In our open-source code, the parameters that govern gRNA selection are explicitly annotated and organized as separate modules. Users who would like to customize or add selection criteria can easily modify the corresponding modules.

Note that beyond positional prioritization, protoSpaceJAM does not score the predicted editing activity of gRNAs. While several algorithms for predicting gRNA activity have been developed, many have been shown to perform inconsistently when applied to experiments using different species, cell lines, or gRNA expression systems (discussed in (33,38)). For example, predicted activity might correlate with the expression levels of gRNA sequences under specific promoter conditions (e.g. using U6 or T7 promoters; see (33)), and not be applicable to experiments that use a different promoter or use direct delivery of purified CRISPR/Cas complexes (39,40). As a consequence, the development of predictive algorithms for gRNA cutting efficiency that generalize across experimental conditions remains an active field of research (38). While we currently do not use activity predictions in our calculations, expert users could easily modify our open-source code for gRNA selection to include an activity prediction score suitable for their experimental context.

Mutational recoding in the HDR donor

protoSpaceJAM supports the optional introduction of silent 'recoding' mutations in two key regions of the HDR donor (Figure 1A; here we use 'recoding' to refer to both silent mutations and mutations in non-coding sequences). The first region is the Cas/gRNA binding site, which may still be present in the homology arm sequences when payload insertion does not destroy the original protospacer. In such cases, knock-in could be impaired if the Cas nuclease either cuts the donor itself during the delivery of reagents in the cell or re-cuts the knock-in allele after DNA repair (16,29,41). This would respectively decrease donor availability or introduce unwanted genomic modifications, negatively impacting knock-in efficiency overall. A well-established practice is therefore to introduce silent mutations to inactivate the gRNA binding site within the HDR donor (16,29,41). protoSpaceJAM uses the Cutting Frequency Determination (CFD) scoring framework established by Doench and colleagues to predict the impact of individual protospacer and PAM mutations on the Cas/gRNA cutting potential for SpCas9 and enAsCas12a (14,42). For each gRNA, protoSpaceJAM identifies the fewest mutations that would bring the maximal CFD score in the donor sequence below a user-defined threshold (default: 0.03). Because the payload sequence itself may by chance contain sequences homologous to the Cas9/gRNA binding site, all positions within the donor (including payload and payload/genome junctions) are considered in this calculation. When recoding within a protein-coding sequence, only silent mutations are used, leveraging maximal sequence divergence between synonymous codons while excluding rare codons (defined by codon usage frequency less than 7.0×10^{-3}, or less than half of the median codon usage for that amino acid). When 'recoding' within a non-coding region, transversion mutations are introduced in up to one of every three bases (Figure 3A). No recoding is allowed in the immediate vicinity of splice junctions, to maintain universally conserved sequence motifs (Supplementary Figure S2). For maximal flexibility, the user can decide whether recoding should prioritize mutations in the PAM region (default) or within the protospacer itself, and optionally restrict the introduction of mutations only to protein-coding regions. While our mutational strategy is designed to always preserve native protein sequences and splice junctions, the introduction of any mutation might affect the function of other non-coding elements (e.g. non-coding RNAs, enhancers, etc.). If preserving non-coding elements is a priority, users can decide to forgo mutational recoding (see below).

Another application of mutational recoding in HDR donors is to increase knock-in efficiency when having to perform Cas/gRNA cuts at a distance from the insertion site. In such cases, introducing silent mutations in the cut-to-insert region prevents the DNA repair tracks from resolving repair before reaching the payload sequence (Supplementary Figure S1A), thereby increasing the rate of payload insertion (16,30). protoSpaceJAM supports recoding within the cut-to-insert region, following the rules outlined above for coding and non-coding sequences and excluding recoding at splice junctions (Figure 3A).

Importantly, the level of recoding for each design is controlled by the user and can be tuned to one of three behaviors: (i) 'no recoding', where no mutations are introduced besides the insertion of the payload itself; (ii) 'prevent re-cut only', where mutations are only introduced to inactivate gRNA binding anywhere in the donor sequence and (iii) 'full recoding', where mutations are introduced to inactivate gRNA binding in the donor, and repair track mutations are introduced in the cut-to-insert region. By default, we recommend using full recoding. The advantage of recoding is illustrated in Figure 3B, C. Figure 3B shows the potential for SpCas9/gRNA re-cutting of knock-in alleles for GFP insertion at the N- or C-terminus of all protein-coding genes in the human, mouse, or zebrafish genomes, as predicted by the maximal residual CFD score within each donor sequence. In the absence of recoding, 53.6%, 55.4% and 61.0% of knock-in sequences exhibit significant re-cutting potential in human, mouse, and zebrafish, respectively (Figure 3B, 'no recoding'; note that when gRNAs proximal to the insertion sites can be chosen, the insertion of the payload alone may be enough to impact the protospacer and prevent re-cutting). The introduction of silent recoding mutations brings the re-cutting potential below the desired threshold in 97.6%, 97.8% and 97.2% of the sequences in human, mouse, and zebrafish, respectively (Figure 3B). In addition, adding repair track mutations in the cut-to-insert region increases knock-in efficiency for gRNAs with large cut-to-insert distances (Figure 3C). To verify this, we measured the integration efficiency of a fluorescent mNeonGreen payload across 32 human protein-coding genes for which top-scoring gRNAs were located at different cut-to-insert distances, comparing designs using the 'full recoding' vs 'prevent re-cut only' modes (Figure 3C, left). The full recoding option led to significantly higher integration efficiency (as measured by the percentage of fluorescent cells detected) for designs in which the cut-to-insert distance was >12 nt (Figure 3C, right panel; note that gene-specific sequence features beyond the cut-to-insert distance are likely to influence knock-in efficiency). These results mirror previous reports in the literature describing the advantages of recoding the cut-to-insert region to increase knock-in efficiency (16,30).

Additional design parameters for dsDNA vs ssODN donors

Double-stranded DNA (dsDNA) and single-stranded oligodeoxynucleotides (ssODNs) are the two main formats for HDR donors (8). ssODNs are widely available through commercial DNA synthesis, are non-toxic, and can be delivered to cells in large amounts to maximize knock-in efficiency (12). However, they are limited to small insertions because their overall length is constrained by the coupling efficiency of chemical DNA synthesis (typically $\leq$ 200 nt). dsDNA donors, which can be delivered to cells in the form of DNA plasmids or linear fragments obtained from PCR amplification or commercial sources (27,43–45), are a more universal form of donor that can be used for insertions of any size. However, they are comparatively more toxic (12) and may require molecular cloning for generating the amounts needed for efficient delivery. Increasingly, sequence-verified dsDNA constructs are becoming commercially available at attractive price points, but commercial dsDNA products are typically limited to 'simple-to-synthesize' sequences with balanced GC content and devoid of homopolymeric stretches. Because homology arms often contain parts of the non-coding genome, HDR donor sequences may often fall beyond the current limitations for commercial dsDNA synthesis.

A key goal of protoSpaceJAM is to provide the user with 'synthesis-ready' donor sequences to streamline the knock-in experimental process. Therefore, the user can choose between two separate donor design modes—dsDNA and ssODN—that use separate design constraints. In dsDNA mode, sequence motifs that might be incompatible with commercial dsDNA synthesis are flagged within the final output table. These flags include homopolymeric runs of 10 or more As and Ts or 6 or more Gs and Cs, and extreme GC content (>65% or

<25% GC content globally or >52% difference in GC content between any given 50 bp stretches). We analyzed the presence of these flags in homology arms of 500 bp flanking the N- and C-terminus tagging sites in all canonical transcripts in human, mouse, and zebrafish genomes (39 637, 43 719 and 55 559 sites, respectively). Our results indicated that homopolymeric runs are the leading impediment to successful synthesis of the dsDNA donor (Figure 4A). Therefore, we incorporated a feature to directly trim the length of homology arms to remove homopolymers from the dsDNA donor (Figure 4B). The user controls whether trimming should be performed, as well as the minimum length of homology arms to be retained. To assist users in making informed decisions regarding the minimum length of homology arms, we analyzed 39 637 N- and C-terminus knock-in sites in the human genome. Our findings indicate that DNA donor arms with lengths of 100, 200 and 500 bp have a 94%, 87% and 70% probability of being free from homopolymers, respectively (Figure 4B). Our trimming algorithm halts once it either meets the minimum length requirement or the sequence becomes free of homopolymers, whichever condition is satisfied first. Trimming can also be used to avoid other user-defined sequence motifs, such as restriction sites for enzymes used in subsequent applications (Figure 4C).

For ssODN synthesis, there is typically no restriction in terms of sequence motifs, but rather in overall length. Therefore, the total length of donors in ssODN mode is capped at a user-defined maximum (default: 200 nt). However, ssODN donors require a choice of polarity for the ssDNA strand to be used. The polarity of the ssODN strand is especially important when using gRNAs with a large cut-to-insert distance, and the most efficient strand orientation depends on whether the cut site is on the 3′ or 5′ side of the insertion site (see Supplementary Figure S1B). This is because ssODN donors predominantly template DNA repair via a synthesis-dependent strand annealing mechanism (30). In this mechanism, 5′-end resection following the Cas-induced double-strand break exposes single-stranded genomic overhangs that can anneal to complementary sequences in the ssODNs donors, which are subsequently extended by DNA synthesis in the 5′ to 3′ direction (Supplementary Figure S1B). This mechanism induces a polarity in the repair process that strongly favors a specific DNA strand for distal payload integration (Supplementary Figure S1B). By default, protoSpaceJAM automatically selects the polarity of the ssODN strand to be in the favored orientation. To validate this design choice, we measured the integration efficiency of a split-mNeonGreen fluorescent payload across 32 human protein-coding genes for which top-scoring gRNAs were located at different cut-to-insert distances, comparing both strand orientations for each ssODN donor (Figure 4D). The favored strand orientation (as defined in Supplementary Figure S1) led to increased integration efficiency, as measured by the percentage of fluorescent cells, particularly for designs in which the cut-to-insert distance was greater than 12 nt (Figure 4D, right panel). Similar strand preference requirements have been previously described and further support our default design choice (15,16,30,41,46). To give the user even finer control over the ssODN strand to be used, four other strand selection modes are also available: Cas/gRNA target vs non-target strand or transcribed vs non-transcribed strand for protein-coding genes and lncRNAs (Figure 4E).

Genotyping primer design for characterization by amplicon-based deep sequencing

After knock-in, the genomic sequence of the edited alleles must be characterized to validate successful editing and the absence of undesired errors introduced during DNA repair. This is often done using deep sequencing of amplicons covering the edit site and homology arms (27,47,48). Depending on the sizes of the knock-in insertion and the DNA donor used, genotyping can be performed using short- or long-read sequencing. Amplification must be driven by primers that bind the genome outside of the homology arms used in the knock-in design (Figure 5), so that amplification cannot be templated from the HDR donor itself but must instead represent bona-fide edited alleles.

We developed a computational pipeline called GenoPrimer that enables rapid primer design for deep sequencing-based genotyping of edited sites (Figure 5A). GenoPrimer is fully integrated into the protoSpaceJAM web tool, allowing users to easily access and utilize this function. Additionally, we provide an open-source standalone version, available at github.com/czbiohub-sf/GenoPrimer. GenoPrimer operates in two distinct modes: short-read and long-read. By default, short-read mode returns primer for 250–350 bp amplicons, and long-read mode returns primers for 3300–3700 bp amplicons, which are designed to span the entire integrated payload (this default size can be changed by the user). Candidate primer pairs are selected using three primary filters: (i) a minimum distance from the edited site of 100 bp for short reads or 1000 bp for long reads. This ensures that at least one primer will bind outside the homology arms, when assuming a combined homology arm length (5′ + 3′) of ≤200 nt for ssODN donors and ≤2000 bp for dsDNA donors; (ii) compliance with default thermodynamic criteria used by Primer3 (24) to ensure optimal PCR efficiency (e.g. similar Tm and low probability of binding and formation of primer dimers and secondary structures). GenoPrimer uses the thermodynamic model from Primer3 to predict Tm. Note that the predicted Tm might differ from the optimal annealing temperature to use with a specific DNA polymerase and (iii) absence of off-target amplicons under 6000 bp, which can arise when both primers in the pair bind non-specifically in opposite orientations on the same off-target chromosome. To identify off-target binding sites, users can select either the Bowtie aligner (25), which is computationally fast but limited to sites with up to three mismatches, or BLAST+ (49), which is slower but allows more mismatches. In cases where no desirable candidates can be found, the default design constraints on ΔTm and amplicon length are iteratively relaxed to nominate alternative primer pairs (Figure 5A). When GenoPrimer successfully identifies suitable primers, it outputs up to three pairs of primers along with their PCR product sizes for unedited alleles and annealing temperatures. We used GenoPrimer to compute genotyping primers using short-read mode for N- and C-terminus insertion sites in all canonical open reading frames in the human, mouse, and zebrafish genomes; primers that fit the selection criteria were successfully identified for 38 190 of 39 200 (97.4%), 41 699 of 43 579 (95.7%) and 48 688 of 50 724 (96.0%) sites, respectively. The same computation using long-read mode yielded primers for 98.4%, 97.6% and 95.0% of sites, respectively. Next, we experimentally tested primers computed for 94 knock-in sites across different human genes (three primer pairs per site). Of these, 93 sites (98.9% of total) yielded primers that fit the selection criteria, with the only exception resulting from an exact sequence duplication between chromosomes that prevented identifying primers devoid of off-target binding sites (CDC26 C-terminus: chr9:113267016–113267441 duplicated on chr7:129409866–129410291). For all 93 sites, at least one primer pair produced an amplicon matching the expected size, as determined by quantitative capillary electrophoresis (Figure 5B). In 96.7% of the cases, the fraction of amplicons of the correct size exceeded 50% of all amplified products, and in 87% of the cases, correct amplicons exceeded 80% of all amplified products (Figure 5C).

Interactive web tool at protospacejam.czbiohub.org

We developed a user-friendly, interactive, flexible, and fast web-based tool to streamline the use of protoSpaceJAM and GenoPrimer (Figure 6). This tool effectively eliminates the need for command-line prompts, enhancing accessibility for users of all skill levels. It utilizes an intuitive step-by-step navigation system, guiding users through each stage of the process – configuration, verification, execution, and download of CRISPR knock-in designs.

Knock-in design jobs on protospacejam.czbiohub.org are built around four separate phases (full manual available at czbiohub-sf.github.io/protoSpaceJAM/index.html). In the '*Build your job*' phase, users can interactively parameterize their design requests online. Alternatively, design features can be input from a user-prepared file through the '*Upload csv*' function. Notably, these two steps can continuously feed the same submission list. This enables different designs using different parameters to be processed in the same job. Once a desired submission list has been populated, that list can be reviewed and confirmed in the '*Review submission*' phase. Next, the design job can be launched in the '*Jam it!*' phase, with the option to also design genotyping primers suitable for either short- or long-read sequencing. Finally, in the '*Results*' phase, users can view gRNAs, donors, and primers (if applicable) for each design and download both the results and the submission list, which includes parameterization information, as csv files (Figure 1B). Annotated HDR donor sequences highlighting the position of homology arms, payload, gRNA sequence, recoding mutations (when applicable) and exon boundaries along with protein-coding translations, can be downloaded in GenBank format.

Users can access documentation, in the style of ReadTheDocs, by following the navigation links situated at the top of the page. For additional assistance, hover tooltips offer immediate contextual explanations of various parameters.

The full source code of our interactive web portal is available at github.com/czbiohub-sf/protoSpaceJAM-portal. Users have the flexibility to set up personalized instances of the web portal, supporting Ensembl-compatible genomes, on their own computing infrastructure.

Validation in human induced pluripotent stem cells (hiPSCs)

Increasingly, human induced pluripotent stem cells (hiPSCs, which can be further differentiated into specific cell types) are being adopted as a powerful system for disease modeling and cell biology research (2,50,51). To validate protoSpaceJAM's ability to design gRNA and HDR donor sequences enabling knock-in in hiPSCs, we measured the integration of a split-GFP fluorescent payload across 34 human protein-coding genes in the WTC11 hiPSC cell line (27). Designs used the single highest-scoring gRNA for each gene, and ssODN HDR donors were generated using the 'full-recoding' mode. GFP-positive cells were detected by flow cytometry for 32 out of 34 edited genes (94% success rate), with a median of 21.7% GFP-positive cells across all experiments (Supplementary Figure S3).

Discussion

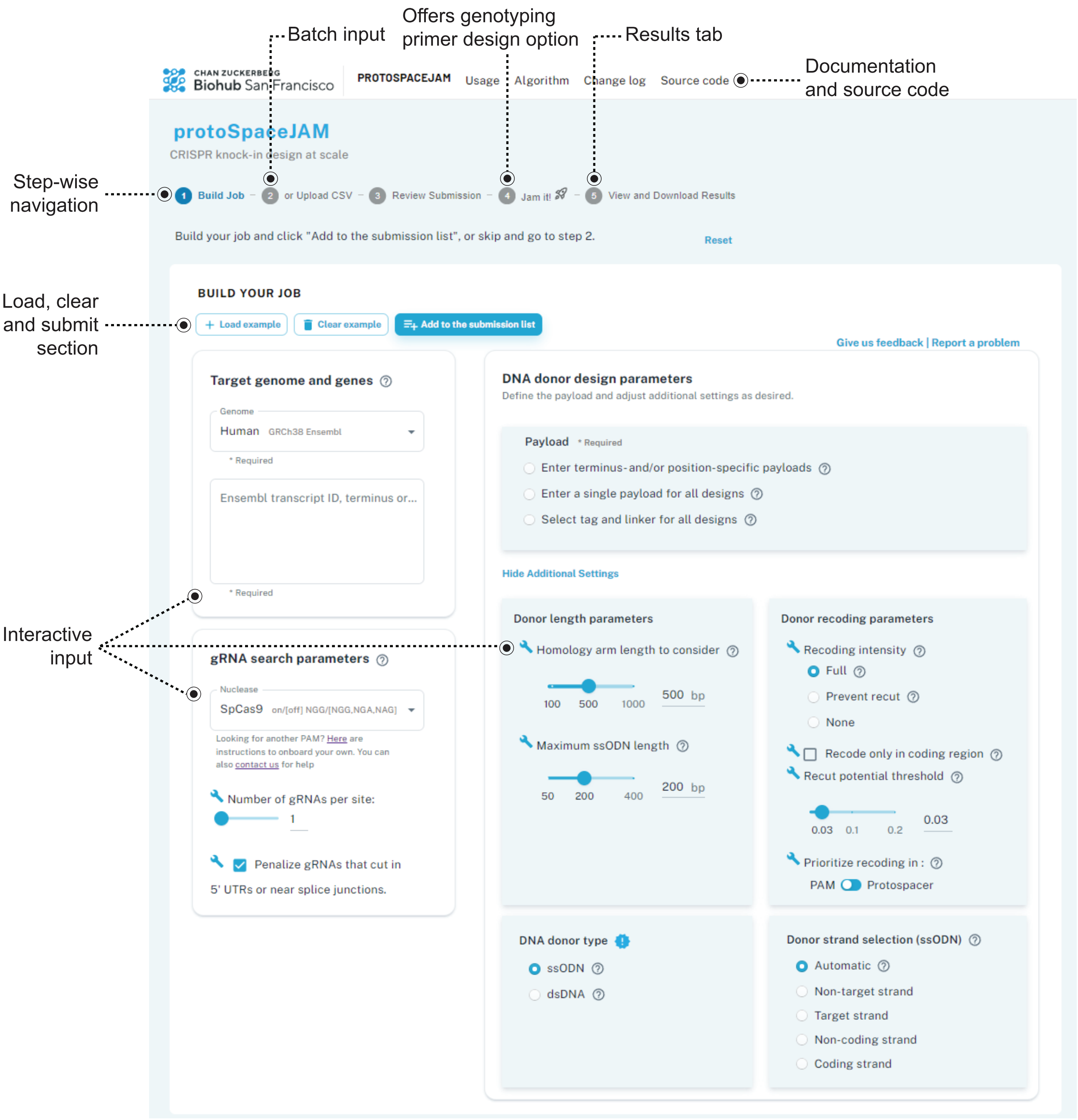
Comparison to existing CRISPR/Cas design tools

We developed protoSpaceJAM to fill an existing gap in the rich ecosystem of CRISPR/Cas design tools. While many algorithms have been developed to select gRNAs for diverse applications (reviewed in Figure 7), few support the design of both gRNA and HDR donors for HDR-based knock-in. In addition, the existing tools for knock-in design are all (to our knowledge) developed by commercial DNA synthesis companies and do not have publicly available source code, limiting their re-use by the community (Figure 7). protoSpaceJAM differentiates itself by being the first fully open-source design tool specifically developed for CRISPR/Cas knock-in. In addition, protoSpaceJAM uses a rich set of explicit and biologically-informed design parameters for gRNA selection and optimal HDR donor construction. In particular, protoSpaceJAM is unique in its ability to (i) penalize gRNAs targeting important regulatory sequences (Figure 2E), (ii) allow users to choose between different levels of HDR donor recoding with silent mutations, including recoding in the cut-to-insert region to increase knock-in efficiency (Figure 3A) and (iii) use different practical design rules for dsDNA versus ssODN donor design, including automated selection of optimal strand orientation for ssODN donors and the design of synthesis-ready sequences for dsDNA donors (Figure 4C, E). In addition, we provide GenoPrimer as a companion algorithm to streamline the design of genotyping primers for deep sequencing using either short or long reads (Figure 5).

Altogether, we have developed an integrated open-source, modular, and fully customizable design platform to accelerate the design and increase the success rate of HDR-based CRISPR/Cas knock-in experiments. The features included in protoSpaceJAM are informed by our own experience in designing over 2000 knock-ins for the OpenCell project (1), in which we systematically characterize the function of human genes by using fluorescent tags to measure the localization and interactions of the corresponding proteins. While we have implemented a sensible set of default design parameters based on the results presented in Figures 2–4, protoSpaceJAM is built to be fully customizable. The web application provides the user with a broad set of feature choices, which can be further tailored to specific applications by modifying the open-source design algorithm. By providing both the standalone open-source code and a user-friendly web interface at protospacejam.czbiohub.org, we aim to enable both expert and novice users to accelerate genetic research for a wide community.

Limitations and future developments

In its current form, protoSpaceJAM is designed for HDR-based, insertional knock-in experiments supporting three Cas enzymes (*SpCas9*, *SpCas9-VQR* and *enAsCas12a*, each targeting different PAMs) in three genomes (human, mouse, and


Batch input
Offers genotyping primer design option
Results tab
Documentation and source code
Step-wise navigation
Load, clear and submit section
Interactive input
CHAN ZUCKERBERG Biohub San Francisco
PROTOSPACEJAM
Usage
Algorithm
Change log
Source code
protoSpaceJAM
CRISPR knock-in design at scale
1 Build Job
2 or Upload CSV
3 Review Submission
4 Jam it!
5 View and Download Results
Reset
+ Load example
Clear example
Add to the submission list
Give us feedback | Report a problem
Hide Additional Settings

Build your job and click "Add to the submission list", or skip and go to step 2.

BUILD YOUR JOB

Target genome and genes

Genome: Human GRCh38 Ensembl
* Required

Ensembl transcript ID, terminus or...
* Required

gRNA search parameters

Nuclease: SpCas9 on/[off] NGG/[NGG,NGA,NAG]

Looking for another PAM? Here are instructions to onboard your own. You can also contact us for help

Number of gRNAs per site: 1

☑ Penalize gRNAs that cut in 5' UTRs or near splice junctions.

DNA donor design parameters

Define the payload and adjust additional settings as desired.

Payload * Required
- ○ Enter terminus-and/or position-specific payloads
- ○ Enter a single payload for all designs
- ○ Select tag and linker for all designs

Donor length parameters

Homology arm length to consider: 500 bp (100, 500, 1000)

Maximum ssODN length: 200 bp (50, 200, 400)

Donor recoding parameters

Recoding intensity
- ● Full
- ○ Prevent recut
- ○ None

□ Recode only in coding region

Recut potential threshold: 0.03 (0.03, 0.1, 0.2)

Prioritize recoding in : PAM / Protospacer

DNA donor type
- ● ssODN
- ○ dsDNA

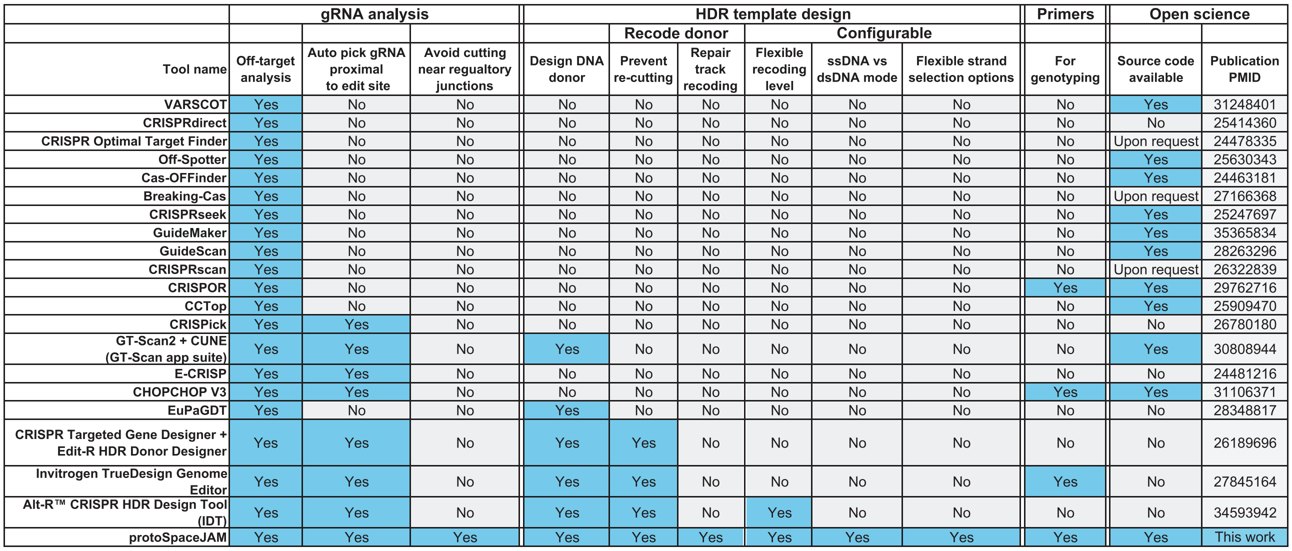
Donor strand selection (ssODN)
- ● Automatic
- ○ Non-target strand
- ○ Target strand
- ○ Non-coding strand
- ○ Coding strand

Figure 6. Annotated screenshot from interactive web tool at protospacejam.czbiohub.org.

zebrafish). The inclusion of additional genomes and Cas proteins with other PAM signatures would further expand the range of our tool (instructions on how to do so are described in our code repository). All protoSpaceJAM designs are currently based on the reference genome sequences for each organism, and do not account for possible sequence polymorphisms present in the genome of a specific cell line or animal strain. In future versions, we plan to enable designs based on user-defined genomic sequence profiles.

Generalizing our algorithm to support knock-in applications for replacement mutagenesis would also broaden its impact and utility, for example to streamline the design of reagents for the introduction of disease-causing polymorphisms. Furthermore, while HDR-based strategies remain the most widely used for knock-in, the development of new approaches such as prime editing is rapidly changing the knock-in method landscape (52). Prime editing bypasses the requirement for double-strand breaks by using a Cas9 nickase fused to a reverse-transcriptase domain, together with a prime editing guide RNA (pegRNA) that bridges the Cas9 gRNA and the insertion payload, which can be integrated following reverse-transcription to its DNA equivalent (7). The design of prime editing experiments would therefore require a new set of constraints, which could be added to protoSpaceJAM.

The current version of GenoPrimer is designed for a genotyping strategy that uses deep sequencing of amplicons spanning the entire genome-integrated payload. Alternative genotyping strategies can also be used, for example using sepa-



	gRNA analysis			HDR template design						Primers	Open science	
					Recode donor		Configurable					
Tool name	Off-target analysis	Auto pick gRNA proximal to edit site	Avoid cutting near regualtory junctions	Design DNA donor	Prevent re-cutting	Repair track recoding	Flexible recoding level	ssDNA vs dsDNA mode	Flexible strand selection options	For genotyping	Source code available	Publication PMID
VARSCOT	Yes	No	No	No	No	No	No	No	No	No	Yes	31248401
CRISPRdirect	Yes	No	No	No	No	No	No	No	No	No	No	25414360
CRISPR Optimal Target Finder	Yes	No	No	No	No	No	No	No	No	No	Upon request	24478335
Off-Spotter	Yes	No	No	No	No	No	No	No	No	No	Yes	25630343
Cas-OFFinder	Yes	No	No	No	No	No	No	No	No	No	Yes	24463181
Breaking-Cas	Yes	No	No	No	No	No	No	No	No	No	Upon request	27166368
CRISPRseek	Yes	No	No	No	No	No	No	No	No	No	Yes	25247697
GuideMaker	Yes	No	No	No	No	No	No	No	No	No	Yes	35365834
GuideScan	Yes	No	No	No	No	No	No	No	No	No	Yes	28263296
CRISPRscan	Yes	No	No	No	No	No	No	No	No	No	Upon request	26322839
CRISPOR	Yes	No	No	No	No	No	No	No	No	Yes	Yes	29762716
CCTop	Yes	No	No	No	No	No	No	No	No	No	Yes	25909470
CRISPick	Yes	Yes	No	No	No	No	No	No	No	No	No	26780180
GT-Scan2 + CUNE (GT-Scan app suite)	Yes	Yes	No	Yes	No	No	No	No	No	No	Yes	30808944
E-CRISP	Yes	Yes	No	No	No	No	No	No	No	No	No	24481216
CHOPCHOP V3	Yes	Yes	No	No	No	No	No	No	No	Yes	Yes	31106371
EuPaGDT	Yes	No	No	Yes	No	No	No	No	No	No	No	28348817
CRISPR Targeted Gene Designer + Edit-R HDR Donor Designer	Yes	Yes	No	Yes	Yes	No	No	No	No	No	No	26189696
Invitrogen TrueDesign Genome Editor	Yes	Yes	No	Yes	Yes	No	No	No	No	Yes	No	27845164
Alt-R™ CRISPR HDR Design Tool (IDT)	Yes	Yes	No	Yes	Yes	No	Yes	No	No	No	No	34593942
protoSpaceJAM	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	This work

Figure 7. Comparison of protoSpaceJAM with published CRISPR/Cas design tools.

rate amplification of each homology arm/genome junction (2). While these alternative strategies are not currently supported by GenoPrimer, its open-source code for '*in silico* PCR' could be adapted to fulfill these applications.

Ultimately, the CRISPR field is characterized by constant innovation, with new methods for genome engineering being developed at a rapid pace. Bioinformatics tools for CRISPR design must therefore be adaptable. The power of open-source tools such as protoSpaceJAM, whose code is written in a modular format, lies in the ease of reuse and adaptation of code elements by anyone in the community, to pave the way for the development of ever-expanding toolsets to propel biological research.

Data availability

protoSpaceJAM and GenoPrimer are implemented in Python and are freely available as both a web server and standalone software under the BSD-3 license. The web server can be accessed at protospacejam.czbiohub.org/ without a login process. The standalone versions can be downloaded from the GitHub repository: github.com/czbiohub-sf/protoSpaceJAM and github.com/czbiohub-sf/GenoPrimer. The internal database of precomputed gRNA information can be downloaded following the usage instructions in the standalone versions. Archived versions of protoSpaceJAM and GenoPrimer (at the time of writing) could be found on Zenodo with DOIs: doi.org/10.5281/zenodo.11057858 and doi.org/10.5281/zenodo.11058148, respectively. Amplicon sequencing data (Figure 2C) were deposited to the Sequence Read Archive database under the identifier PRJNA1026361.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank J. Hanks, C. Kapfer and R. Alvarez for help with computational resource support. Some figures were created with BioRender.com. We thank S. Schmid for critical feedback. We thank our donors, Priscilla Chan and Mark Zuckerberg, for their generous support.

Author contributions: Duo Peng: Conceptualization, Formal analysis, Methodology, Validation, Writing—original draft, review & editing. Madhuri Vangipuram: Formal analysis, Validation. Joan Wong: Conceptualization, Writing—review & editing. Manuel Leonetti: Conceptualization, Formal analysis, Methodology, Validation, Writing—original draft, review & editing.

Funding

Non-profit research institution Chan Zuckerberg Biohub San Francisco in San Francisco, California, USA. Funding for open access charge: Chan Zuckerberg Biohub 501 c3 non-profit.

Conflict of interest statement

None declared.

References

1. 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*, **375**, eabi6983.
2. Roberts,B., Haupt,A., Tucker,A., Grancharova,T., Arakaki,J., Fuqua,M.A., Nelson,A., Hookway,C., Ludmann,S.A., Mueller,I.A., *et al.* (2017) Systematic gene tagging using CRISPR/Cas9 in human stem cells to illuminate cell organization. *Mol. Biol. Cell*, **28**, 2854–2874.
3. Ramos,D.M., Skarnes,W.C., Singleton,A.B., Cookson,M.R. and Ward,M.E. (2021) Tackling neurodegenerative diseases with genomic engineering: a new stem cell initiative from the NIH. *Neuron*, **109**, 1080–1083.
4. Shy,B.R., Vykunta,V.S., Ha,A., Talbot,A., Roth,T.L., Nguyen,D.N., Pfeifer,W.G., Chen,Y.Y., Blaeschke,F., Shifrut,E., *et al.* (2023) High-yield genome engineering in primary cells using a hybrid ssDNA repair template and small-molecule cocktails. *Nat. Biotechnol.*, **41**, 521–531.
5. Fichter,K.M., Setayesh,T. and Malik,P. (2023) Strategies for precise gene edits in mammalian cells. *Mol. Ther. Nucleic Acids*, **32**, 536–552.

6. Suzuki,K., Tsunekawa,Y., Hernandez-Benitez,R., Wu,J., Zhu,J., Kim,E.J., Hatanaka,F., Yamamoto,M., Araoka,T., Li,Z., *et al.* (2016) In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. *Nature*, **540**, 144–149.
7. Anzalone,A.V., Randolph,P.B., Davis,J.R., Sousa,A.A., Koblan,L.W., Levy,J.M., Chen,P.J., Wilson,C., Newby,G.A., Raguram,A., *et al.* (2019) Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, **576**, 149–157.
8. Lau,C.-H., Tin,C. and Suh,Y. (2020) CRISPR-based strategies for targeted transgene knock-in and gene correction. *Fac. Rev.*, **9**, 20.
9. Xu,X., Gao,D., Wang,P., Chen,J., Ruan,J., Xu,J. and Xia,X. (2018) Efficient homology-directed gene editing by CRISPR/Cas9 in human stem and primary cells using tube electroporation. *Sci. Rep.*, **8**, 11649.
10. Dewari,P.S., Southgate,B., Mccarten,K., Monogarov,G., O'Duibhir,E., Quinn,N., Tyrer,A., Leitner,M.-C., Plumb,C., Kalantzaki,M., *et al.* (2018) An efficient and scalable pipeline for epitope tagging in mammalian stem cells using Cas9 ribonucleoprotein. *eLife*, **7**, e35069.
11. Skarnes,W.C., Pellegrino,E. and McDonough,J.A. (2019) Improving homology-directed repair efficiency in human stem cells. *Methods*, **164–165**, 18–28.
12. Roth,T.L., Puig-Saus,C., Yu,R., Shifrut,E., Carnevale,J., Li,P.J., Hiatt,J., Saco,J., Krystofinski,P., Li,H., *et al.* (2018) Reprogramming human T cell function and specificity with non-viral genome targeting. *Nature*, **559**, 405–409.
13. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O., *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.
14. 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.
15. Liang,X., Potter,J., Kumar,S., Ravinder,N. and Chesnut,J.D. (2017) Enhanced CRISPR/Cas9-mediated precise genome editing by improved design and delivery of gRNA, Cas9 nuclease, and donor DNA. *J. Biotechnol.*, **241**, 136–146.
16. Schubert,M.S., Thommandru,B., Woodley,J., Turk,R., Yan,S., Kurgan,G., McNeill,M.S. and Rettig,G.R. (2021) Optimized design parameters for CRISPR Cas9 and Cas12a homology-directed repair. *Sci. Rep.*, **11**, 19482.
17. Richardson,C.D., Ray,G.J., DeWitt,M.A., Curie,G.L. and Corn,J.E. (2016) Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat. Biotechnol.*, **34**, 339–344.
18. Wang,Y., Liu,K.I., Sutrisnoh,N.-A.B., Srinivasan,H., Zhang,J., Li,J., Zhang,F., Lalith,C.R.J., Xing,H., Shanmugam,R., *et al.* (2018) Systematic evaluation of CRISPR-Cas systems reveals design principles for genome editing in human cells. *Genome Biol.*, **19**, 62.
19. Kleinstiver,B.P., Prew,M.S., Tsai,S.Q., Topkar,V.V., Nguyen,N.T., Zheng,Z., Gonzales,A.P.W., Li,Z., Peterson,R.T., Yeh,J.-R.J., *et al.* (2015) Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, **523**, 481–485.
20. Kleinstiver,B.P., Sousa,A.A., Walton,R.T., Tak,Y.E., Hsu,J.Y., Clement,K., Welch,M.M., Horng,J.E., Malagon-Lopez,J., Scarfò,I., *et al.* (2019) Engineered CRISPR-Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. *Nat. Biotechnol.*, **37**, 276–282.
21. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
22. Concordet,J.-P. and Haeussler,M. (2018) CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.*, **46**, W242–W245.
23. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
24. Untergasser,A., Cutcutache,I., Koressaar,T., Ye,J., Faircloth,B.C., Remm,M. and Rozen,S.G. (2012) Primer3–new capabilities and interfaces. *Nucleic Acids Res.*, **40**, e115.
25. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
26. Ye,J., Coulouris,G., Zaretskaya,I., Cutcutache,I., Rozen,S. and Madden,T.L. (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinf.*, **13**, 134.
27. Canaj,H., Hussmann,J.A., Li,H., Beckman,K.A., Goodrich,L., Cho,N.H., Li,Y.J., Santos,D.A., McGeever,A., Stewart,E.M., *et al.* (2019) Deep profiling reveals substantial heterogeneity of integration outcomes in CRISPR knock-in experiments. bioRxiv doi: https://doi.org/10.1101/841098, 13 November 2019, preprint: not peer reviewed.
28. Clement,K., Rees,H., Canver,M.C., Gehrke,J.M., Farouni,R., Hsu,J.Y., Cole,M.A., Liu,D.R., Joung,J.K., Bauer,D.E., *et al.* (2019) CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.*, **37**, 224–226.
29. Paquet,D., Kwart,D., Chen,A., Sproul,A., Jacob,S., Teo,S., Olsen,K.M., Gregg,A., Noggle,S. and Tessier-Lavigne,M. (2016) Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature*, **533**, 125–129.
30. Paix,A., Folkmann,A., Goldman,D.H., Kulaga,H., Grzelak,M.J., Rasoloson,D., Paidemarry,S., Green,R., Reed,R.R. and Seydoux,G. (2017) Precision genome editing using synthesis-dependent repair of Cas9-induced DNA breaks. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E10745–E10754.
31. Lin,Y., Cradick,T.J., Brown,M.T., Deshmukh,H., Ranjan,P., Sarode,N., Wile,B.M., Vertino,P.M., Stewart,F.J. and Bao,G. (2014) CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.*, **42**, 7473–7485.
32. Cho,S.W., Kim,S., Kim,Y., Kweon,J., Kim,H.S., Bae,S. and Kim,J.-S. (2014) Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.*, **24**, 132–141.
33. Haeussler,M., Schönig,K., Eckert,H., Eschstruth,A., Mianné,J., Renaud,J.-B., Schneider-Maunoury,S., Shkumatava,A., Teboul,L., Kent,J., *et al.* (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.*, **17**, 148.
34. O'Brien,A.R., Wilson,L.O.W., Burgio,G. and Bauer,D.C. (2019) Unlocking HDR-mediated nucleotide editing by identifying high-efficiency target sites using machine learning. *Sci. Rep.*, **9**, 2788.
35. Bialk,P., Rivera-Torres,N., Strouse,B. and Kmiec,E.B. (2015) Regulation of gene editing activity directed by single-stranded oligonucleotides and CRISPR/Cas9 systems. *PLoS One*, **10**, e0129308.
36. Kan,Y., Ruis,B., Takasugi,T. and Hendrickson,E.A. (2017) Mechanisms of precise genome editing using oligonucleotide donors. *Genome Res.*, **27**, 1099–1111.
37. Hinnebusch,A.G., Ivanov,I.P. and Sonenberg,N. (2016) Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science*, **352**, 1413–1416.
38. Konstantakos,V., Nentidis,A., Krithara,A. and Paliouras,G. (2022) CRISPR-Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Res.*, **50**, 3616–3637.
39. Kim,S., Kim,D., Cho,S.W., Kim,J. and Kim,J.-S. (2014) Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.*, **24**, 1012–1019.
40. Lin,S., Staahl,B.T., Alla,R.K. and Doudna,J.A. (2014) Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife*, **3**, e04766.

41. Okamoto,S., Amaishi,Y., Maki,I., Enoki,T. and Mineno,J. (2019) Highly efficient genome editing for single-base substitutions using optimized ssODNs with Cas9-RNPs. *Sci. Rep.*, **9**, 4811.
42. DeWeirdt,P.C., Sanson,K.R., Sangree,A.K., Hegde,M., Hanna,R.E., Feeley,M.N., Griffith,A.L., Teng,T., Borys,S.M., Strand,C., *et al.* (2021) Optimization of AsCas12a for combinatorial genetic screens in human cells. *Nat. Biotechnol.*, **39**, 94–104.
43. Ghanta,K.S., Chen,Z., Mir,A., Dokshin,G.A., Krishnamurthy,P.M., Yoon,Y., Gallant,J., Xu,P., Zhang,X.-O., Ozturk,A.R., *et al.* (2021) 5'-Modifications improve potency and efficacy of DNA donors for precision genome editing. *eLife*, **10**, e72216.
44. Kanca,O., Zirin,J., Garcia-Marques,J., Knight,S.M., Yang-Zhou,D., Amador,G., Chung,H., Zuo,Z., Ma,L., He,Y., *et al.* (2019) An efficient CRISPR-based strategy to insert small and large fragments of DNA using short homology arms. *eLife*, **8**, e51539.
45. 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.
46. Harmsen,T., Klaasen,S., van de Vrugt,H. and Te Riele,H. (2018) DNA mismatch repair and oligonucleotide end-protection promote base-pair substitution distal from a CRISPR/Cas9-induced DNA break. *Nucleic Acids Res.*, **46**, 2945–2955.
47. Ran,F.A., Hsu,P.D., Lin,C.-Y., Gootenberg,J.S., Konermann,S., Trevino,A.E., Scott,D.A., Inoue,A., Matoba,S., Zhang,Y., *et al.* (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, **154**, 1380–1389.
48. Hendel,A., Kildebeck,E.J., Fine,E.J., Clark,J., Punjya,N., Sebastiano,V., Bao,G. and Porteus,M.H. (2014) Quantifying genome-editing outcomes at endogenous loci with SMRT sequencing. *Cell Rep.*, **7**, 293–305.
49. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinf.*, **10**, 421.
50. Karagiannis,P., Takahashi,K., Saito,M., Yoshida,Y., Okita,K., Watanabe,A., Inoue,H., Yamashita,J.K., Todani,M., Nakagawa,M., *et al.* (2019) Induced pluripotent stem cells and their use in Human models of disease and development. *Physiol. Rev.*, **99**, 79–114.
51. Pantazis,C.B., Yang,A., Lara,E., McDonough,J.A., Blauwendraat,C., Peng,L., Oguro,H., Kanaujiya,J., Zou,J., Sebesta,D., *et al.* (2022) A reference human induced pluripotent stem cell line for large-scale collaborative studies. *Cell Stem Cell*, **29**, 1685–1702.
52. Anzalone,A.V., Koblan,L.W. and Liu,D.R. (2020) Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.*, **38**, 824–844.

Received: October 13, 2023. **Revised:** June 6, 2024. **Editorial Decision:** June 11, 2024. **Accepted:** June 17, 2024