nature biotechnology

Article

https://doi.org/10.1038/s41587-022-01527-4

# Drag-and-drop genome insertion of large sequences without double-strand DNA cleavage using CRISPR-directed integrases

Received: 30 April 2022

Accepted: 23 September 2022

Published online: 24 November 2022

Check for updates

Matthew T. N. Yarnall[1,11], Eleonora I. Ioannidi[1,2,11], Cian Schmitt-Ulms[1,11], Rohan N. Krajeski[1,11], Justin Lim[1], Lukas Villiger[1], Wenyuan Zhou[1], Kaiyi Jiang[1,3], Sofya K. Garushyants[4], Nathaniel Roberts[5], Liyang Zhang[5], Christopher A. Vakulskas[5], John A. Walker II[6], Anastasia P. Kadina[6], Adrianna E. Zepeda[6], Kevin Holden[6], Hong Ma[7], Jun Xie[7], Guangping Gao[7], Lander Foquet[8], Greg Bial[8], Sara K. Donnelly[9], Yoshinari Miyata[9], Daniel R. Radiloff[9], Jordana M. Henderson[10], Andrew Ujita[10], Omar O. Abudayyeh[1,12] ✉ & Jonathan S. Gootenberg[1,12] ✉

Programmable genome integration of large, diverse DNA cargo without DNA repair of exposed DNA double-strand breaks remains an unsolved challenge in genome editing. We present programmable addition via site-specific targeting elements (PASTE), which uses a CRISPR–Cas9 nickase fused to both a reverse transcriptase and serine integrase for targeted genomic recruitment and integration of desired payloads. We demonstrate integration of sequences as large as ~36 kilobases at multiple genomic loci across three human cell lines, primary T cells and non-dividing primary human hepatocytes. To augment PASTE, we discovered 25,614 serine integrases and cognate attachment sites from metagenomes and engineered orthologs with higher activity and shorter recognition sequences for efficient programmable integration. PASTE has editing efficiencies similar to or exceeding those of homology-directed repair and non-homologous end joining-based methods, with activity in non-dividing cells and in vivo with fewer detectable off-target events. PASTE expands the capabilities of genome editing by allowing large, multiplexed gene insertion without reliance on DNA repair pathways.

Programmable genome insertion is vital for both gene therapy and basic research. Common methods to insert long DNA sequences rely on cellular responses to double-strand breaks (DSBs) using programmable nucleases, such as CRISPR–Cas9[1–3], for induction of repair pathways such as non-homologous end joining (NHEJ)[4], as with the homology-independent targeted insertion (HITI)[5] technology or homology-directed repair (HDR)[6–8]. However, DSB-based approaches have limitations. Genome damage causes undesirable outcomes, including insertions/deletions

[1] McGovern Institute for Brain Research at MIT, Massachusetts Institute of Technology, Cambridge, MA, USA. [2] ETH Zürich, Zürich, Switzerland. [3] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [4] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. [5] Integrated DNA Technologies, Coralville, IA, USA. [6] Synthego Corporation, Redwood City, CA, USA. [7] University of Massachusetts Chan Medical School, Worcester, MA, USA. [8] Yecuris Corporation, Tualatin, OR, USA. [9] PhoenixBio USA Corporation, New York, NY, USA. [10] TriLink Biotechnologies LLC, San Diego, CA, USA. [11] These authors contributed equally: Matthew T. N. Yarnall, Eleonora I. Ioannidi, Cian Schmitt-Ulms, Rohan N. Krajeski. [12] These authors jointly supervised this work: Omar O. Abudayyeh, Jonathan S. Gootenberg. ✉ e-mail: omarabu@mit.edu; jgoot@mit.edu

(indels), translocations and activation of p53 (refs. [9,10]). NHEJ can generate off-target insertions at unintended DSBs[11], and HDR has low efficiency in non-dividing cells, including many cell types in vivo, and requires long DNA templates that are labor-intensive to produce[12]. Genome-editing technologies such as base editing[13–15] and prime editing[16] alleviate DSB dependencies but are limited to only nucleotide edits, small insertions (less than ~50 nucleotides) or short deletions (less than ~80 nucleotides)[16] and cannot install or replace large sequences of DNA. More recent paired guide prime-editing approaches, which use two prime-editing guide RNAs (pegRNAs) with complementary reverse transcription template regions, have enabled insertion of large sequences by biasing repair toward the edited strands[17,18]. However, these approaches have diminishing efficiency in the 1- to 5.6-kilobase (kb) range and cannot insert larger sequences.

Natural transposable element systems, which include several families of integrases and transposases, provide efficient routes for genome integration without DSBs but lack the programmability of CRISPR effector nucleases. Transposases insert varying copies of a donor sequence into cells at loosely defined sites, such as TA dinucleotides, resulting in semirandom gene insertion throughout the genome[19]. By contrast, site-specific integrases, such as large serine phage integrases, efficiently integrate DNA cargo into sequence-defined landing sites that are ~30–50 nucleotides long[20] and have been used to insert therapeutic transgenes at naturally occurring pseudosites in the human genome in preclinical models[21]. While targeted integration can be achieved by a two-step approach involving prior insertion of integrase landing sites at a desired location using HDR[22], this approach is limited in efficiency and the risks associated with DSBs. Furthermore, a major issue limiting clinical application of certain integrases, such as phiC31, is chromosomal rearrangements between pseudosites, which can also lead to major DNA damage responses[23,24].

Engineered systems to direct integrases, recombinases or transposases to genomic sites for integration of gene cargos without DNA cleavage rely on fusions with programmable DNA-binding proteins. Approaches fusing zinc fingers, transcription activator-like effectors or catalytically inactive Cas9 programmable DNA-binding proteins to transposases[25–29] or recombinases[30–36] have been demonstrated in mammalian cells, but their reported integration efficiency is low at genomic loci. Moreover, transposase fusions are hindered by excessive promiscuity and off-target insertions, while recombinase fusions have limited targets in the genome due to intrinsic sequence restrictions.

To overcome the current limitations of gene integration approaches, we married advances in programmable CRISPR-based gene editing, such as prime editing, with precise site-specific integrases. Fusing Cas9, reverse transcriptases and large serine integrases, we demonstrate programmable integration of cargos up to ~36 kb in a single delivery reaction with efficiencies up to ~50–60% in cell lines and ~4–5% in primary human hepatocytes and T cells. This approach, termed programmable addition via site-specific targeting elements (PASTE), is easily retargeted to new genes, can be delivered with a single dose of plasmids, and functions in non-dividing and primary cells. By profiling thousands of guide designs in a pooled screen, we determined guide rules for optimal programming to loci. We engineered PASTE for orthogonal integration by simultaneously introducing three genes at three separate loci, and sequence replacement by concurrently deleted and inserted sequences using guide pairs. With genome-wide sequencing, we show that PASTE is much more specific than HITI, with higher insertion purity than HDR and HITI. Comparing PASTE to other prime-editing-based insertion approaches[17], we found 8.3- to 42.1-fold higher integration efficiencies by PASTE at three endogenous targets. To further improve PASTE, we mined bacterial genomes and metagenomes for integrases, found 25,614 new integrase orthologs and predicted associated attachment sites, demonstrated activity of select recombinase orthologs in mammalian cells and used these integrases for high-efficiency integration as part of the PASTE system. For therapeutic relevance, we show that diverse templates are compatible with PASTE, including adenovirus-associated virus (AAV) and adenovirus (AdV), allowing for DNA integration of viruses and other DNA templates and extend our use of PASTE into mouse models for in vivo programmable gene insertion in the liver. As a genome editing tool, PASTE opens multiple applications for gene insertion and tagging in biomedical research and therapeutic development.

## Results

### PASTE combines CRISPR editing and site-specific integration

We envisioned a programmable integration system coupling a CRISPR-based targeting approach with efficient insertion via serine integrases, which typically insert sequences containing an *attP* attachment site into a target containing the related *attB* attachment site. By using programmable genome editing to place integrase landing sites at desired locations in the genome, this system would guide the direct activity of the associated integrase to the specific genomic site. As prime editors have been reported to insert 44-base pair (bp) sequences[16], we hypothesized that the ~46-bp *attB* landing site of serine integrases could be incorporated into the pegRNA design and copied into the genome via reverse transcription and flap repair (Fig. 1a,b). This 'beacon' would serve as a target for an integrase, which could either be supplied in *trans* or directly fused to the Cas9 protein for additional recruitment. By simultaneously delivering a circular double-stranded DNA template containing the *attP* attachment site, the expressed integrase could directly integrate the DNA cargo at the desired target site with a single delivery mechanism (Fig. 1a,b).

We engineered pegRNAs with *attB* sequences, hereafter referred to as attachment site-containing guide RNA (atgRNA), and surveyed a panel of atgRNAs with different length *attB* truncations, successfully inserting sequences up to 56 bp at the β-actin (*ACTB*) gene locus, with higher efficiency at lengths below 31 bp (Extended Data Fig. 1a,b). As prime editing has been reported to insert *loxP* beacons for Cre-based insertion[16], we tested a Cre-based integration approach with coexpression of PE2 and a Cre recombinase; however, tyrosine recombinases showed inefficient insertion (Fig. 1c). Given the high efficiency of serine recombinases[37], we evaluated a panel of multiple enzymes, including Bxb1 (hereafter referred to as BxbINT), TP901 (hereafter referred to as Tp9INT) and phiBT1 (hereafter referred to as Bt1INT) phage serine integrases (Supplementary Table 1) and could insert all landing sites tested, with efficiencies between 10 and 30% (Extended Data Fig. 1c). To test the complete system, we combined all components and delivered them in a single transfection: the prime-editing vector, the atgRNA, a nicking guide for stimulating repair of the other strand, a mammalian expression vector for the corresponding integrase or recombinase and a 969-bp minicircle[38] DNA cargo encoding green fluorescent protein (GFP; Fig. 1d). We compared GFP integration rates among the four integrases and recombinases and found that BxbINT integrase showed the highest integration rate (~15%) at the targeted *ACTB* locus and required the nicking guide for optimal performance (Fig. 1d and Extended Data Fig. 1d–f). This combined system, termed PASTEv1, resulted in programmable efficient insertion of the enhanced GFP (EGFP) transgene.

We next hypothesized that we could improve PASTE editing through a series of protein and guide engineering efforts. We tested modified scaffold designs (atgRNAv2) for increased stabilization and expression from RNA polymerase III promoters[39], improving both atgRNA landing site insertion and overall PASTE efficiency (Extended Data Fig. 1g). To optimize other potential bottlenecks for PASTE activity, we screened a panel of protein modifications at the *ACTB* and lamin B1 (*LMNB1*) loci, including alternative reverse transcriptase fusions and mutations, various linkers between the Cas9, reverse transcriptase and integrase domains and reverse transcriptase and BxbINT domain mutants (Fig. 1e, Extended Data Fig. 1h–k and Supplementary Tables 2 and 3). Several protein modifications, including a 48-residue XTEN linker between the Cas9 and reverse transcriptase, and the fusion of MMuLV to the Sto7d DNA-binding domain or mutation of L139P[40]

https://doi.org/10.1038/s41587-022-01527-4

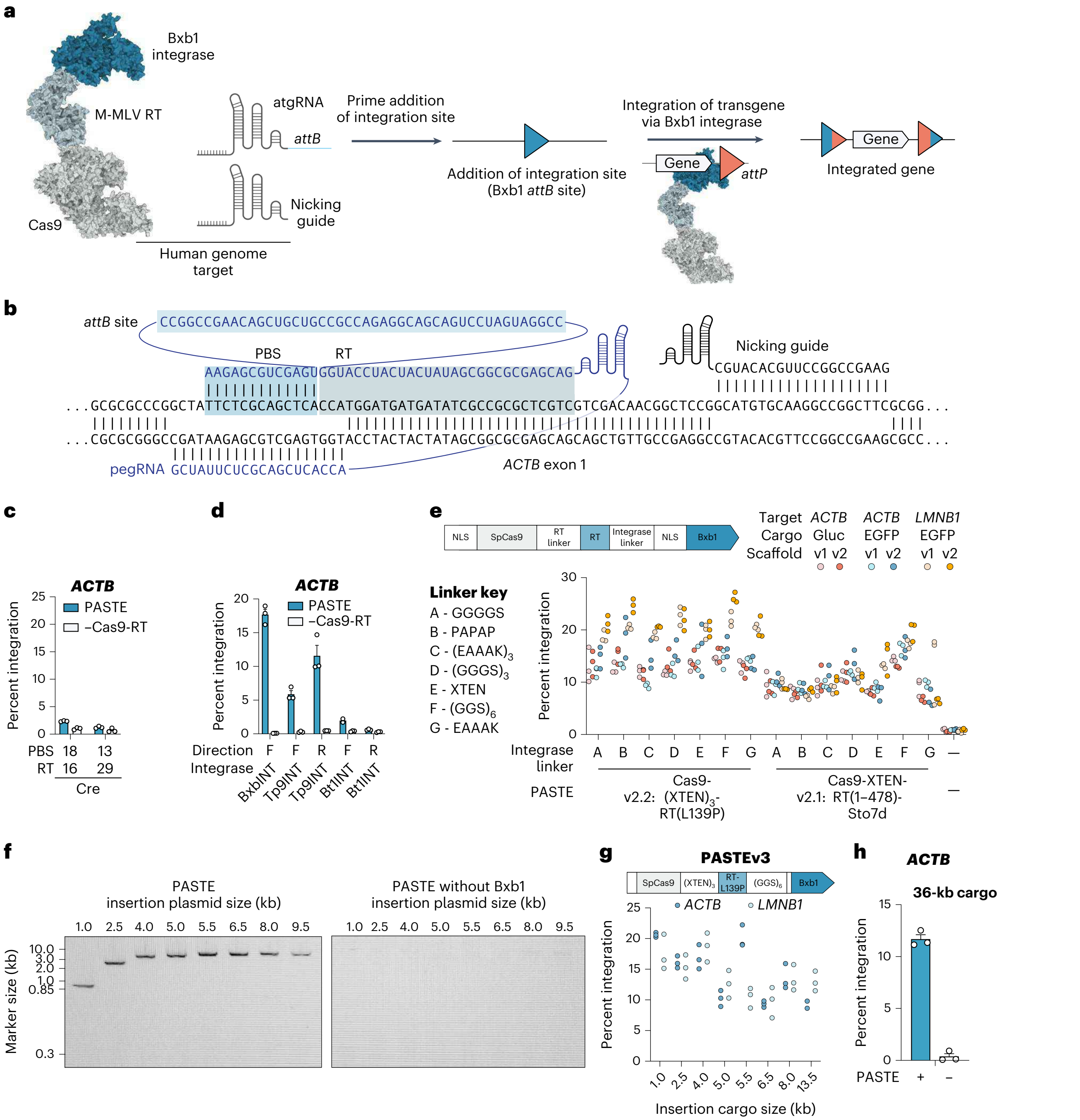


**Fig. 1 | PASTE editing allows for programmable gene insertion independent of DNA repair pathways. a**, Schematic of programmable gene insertion with PASTE. The PASTE system involves insertion of landing sites via Cas9-directed reverse transcriptases, followed by landing site recognition and integration of cargo via Cas9-directed integrases. **b**, Schematic of PASTE insertion at the *ACTB* locus, showing guide and target sequences. **c**, Comparison of GFP cargo integration efficiency between BxbINT and Cre recombinase at the 5′ end of the *ACTB* locus. **d**, Comparison of PASTE integration efficiency of GFP with a panel of integrases targeting the 5′ end of the *ACTB* locus. Both orientations of landing sites are profiled (F, forward; R, reverse). **e**, Optimization of PASTE constructs with a panel of linkers and RT modifications for EGFP integration at the *ACTB* and *LMNB1* loci with different payloads; NLS, nuclear localization sequence. **f**, Gel electrophoresis showing complete insertion by PASTE for multiple cargo sizes. **g**, Effect of cargo size on PASTEv3 insertion efficiency at the endogenous *ACTB* and *LMNB1* targets. Cargos were transfected with fixed molar amounts. **h**, PASTEv3 insertion of a 36-kb cargo template at the *ACTB* locus. Data are shown as mean ± s.e.m.; $n = 3$.

improved PASTE integration efficiency (Extended Data Fig. 1h–j). When these top modifications were combined with a $(GGS)_6$ linker between the reverse transcriptase and BxbINT, they produced up to ~30% gene integration, highlighting the importance of directly recruiting the integrase to the target site (Fig. 1e and Extended Data Fig. 1k). We refer to this optimized construct, SpCas9-(XTEN-48)-RT(L139P)-$(GGS)_6$-BxbINT, as PASTEv2. We combined PASTEv2 with atgRNAv2 to generate PASTEv3, which achieved precise integration of templates as large as ~36,000 bp with ~10–20% integration efficiency at *ACTB* and *LMNB1* (Fig. 1f–h and Extended Data Fig. 2a–e), with complete integration of the full-length cargo confirmed by Sanger sequencing (Extended Data Fig. 2f,g).

## atgRNA and *attB* site parameters influence PASTE efficiency

To optimize atgRNA parameters for PASTE, we explored the impact of atgRNA and integrase parameters on integration efficiency. Relevant atgRNA parameters for PASTEv1 include the primer binding site (PBS), reverse transcription template (RT) and *attB* site lengths and the relative locations and efficacy of the atgRNA spacer and nicking guide (Extended Data Fig. 3a). We tested a range of PBS and RT lengths at *ACTB* and *LMNB1* and found that rules governing integration efficiency varied between loci, with shorter PBS lengths and longer RT designs having higher integration rates at the *ACTB* locus (Extended Data Fig. 3b) and longer PBS and shorter RT designs performing better at *LMNB1* (Extended Data Fig. 3c). These differences may be related to locus-dependent efficiency of priming and resolution of flap insertion observed in other prime-editing applications[16]. The length of the *attB* landing site must balance two conflicting factors: the higher efficiency of prime editing for smaller inserts[16] and reduced efficiency of Bxb1 integration at shorter *attB* lengths[41]. We evaluated *attB* lengths at *ACTB*, *LMNB1* and nucleolar phosphoprotein p130 (*NOLC1*) loci and found that the optimal *attB* length was locus dependent. At the *ACTB* locus, long *attB* lengths could be inserted (Extended Data Fig. 1a), and overall PASTE efficiencies for the insertion of GFP were highest for long *attB* lengths (Extended Data Fig. 3d). By contrast, intermediate *attB* lengths had higher overall integration efficiencies (>20%) at *LMNB1* (Extended Data Fig. 3e) and *NOLC1* (Extended Data Fig. 3f), indicating that the increased efficiency of installing shorter *attB* sequences overcame the reduction of BxbINT integration at these sites. We tested a panel of shorter RT and PBS guides at *ACTB* and *LMNB1* loci in comparison to our previous optimized guides and found that while shorter RT and PBS sequences did not increase integration at *ACTB* (Extended Data Fig. 3g), they had improved integration at *LMNB1* (Extended Data Fig. 3h). Moreover, manual design of a variety of atgRNAs to different targets had varying levels of performance and integration outcomes at seven different gene loci (*ACTB*, *SUPT16H*, *SRRM2*, *NOLC1*, *DEPDC4*, *NES* and *LMNB1*; Extended Data Fig. 3i).

To develop thorough rules for design, we tested atgRNA designs in high throughput via pooled library screening (Fig. 2a). Using pooled oligonucleotide synthesis and cloning, we generated a library of 10,580 atgRNA designs for 11 spacers across 8 target genes (*ACTB*, *LMNB1*, *NOLC1*, *SUPT16H*, *DEPDC4*, *NES*, *CFTR* and *SERPINA1*). For each spacer/target pair, we were able to evaluate PBS lengths between 5 and 19 bp, RT lengths between 6 and 36 bp (increments of two bases) and *attB* lengths of 38, 40, 43 and 46 bp, generating a distribution of edits (Fig. 2b and Supplementary Data 1 and 2). Across the screen, every gene had atgRNAs with significant *attB* insertion rates (Fig. 2b,c). After analyzing the results, we found that more *attB* insertion was generally found at a per-target basis for shorter *attB* sites and that a wider range of RT and PBS lengths was permissible, although the exact optimal combinations differed across genes (Fig. 2d and Extended Data Fig. 4). Across the eight targets, RTs longer than 20 bp tended to yield higher *attB* insertion rates, whereas PBS lengths could be between 5 and 19 bp without any clear trend. To validate the screen, we tested a panel of top-predicted atgRNAs and found that they were all capable of higher-efficiency *attB* insertion (Fig. 2e) and PASTE integration (Fig. 2f) than our previous set of manually designed atgRNAs derived from our arrayed screening of parameters.

To build an explicit predictive model for designing atgRNAs for PASTE, we trained a classifier using a *k*-mer-based multilayer perceptron (MLP) for modeling the effect of an atgRNA sequence on the final editing rate of *attB* insertion. Feature optimization and model training had high accuracy (area under the curve (AUC) = 0.84; Fig. 2g), and scoring of atgRNAs not seen by the model against *LMNB1*, *NOLC1* and *ACTB* revealed clear differences in efficiency between guides nominated by the model and those rejected (Fig. 2h,i). Because our screening results have shown that rational design rules are difficult to generalize across gene targets, we released this prediction model as a guide design tool via a software package (https://github.com/abugoot-lab/pegRNA_rank) that simply receives as input a user's target sequence and produces a list of atgRNAs rank ordered by the predicted efficiency score.

The PE3 version of prime editing combines PE2 and an additional nicking guide to bias resolution of the flap intermediate toward insertion. To test the importance of nicking guide selection on PASTE editing, we tested integration at *ACTB* and *LMNB1* loci with two nicking guide positions. Suboptimal nicking guide positions reduced PASTE efficiency up to 30% (Extended Data Fig. 5a,b), in agreement with the 75% reduction of PASTE efficiency in the absence of nicking guide (Extended Data Figs. 1d and 5c). We also found, as expected, that the atgRNA spacer sequence was necessary for PASTE integration, and substitution of the spacer sequence with a non-targeting guide eliminated editing (Extended Data Fig. 5d).

## PASTE integration with diverse payloads at endogenous sites

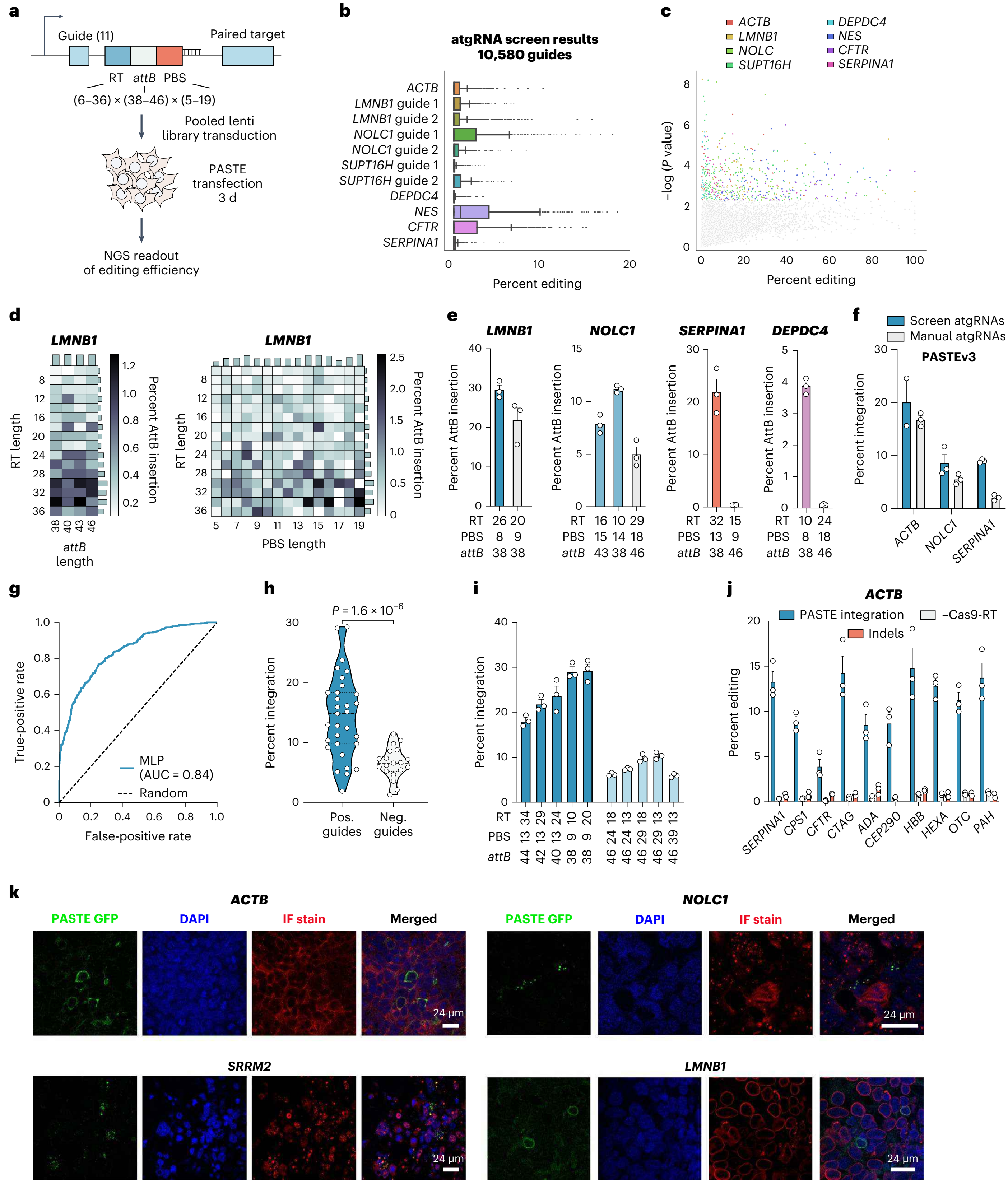
Because PASTE does not require locus homology on cargo plasmids, integration of diverse cargo sequences is modular and easily scaled across different loci. With PASTEv3, we tested a panel of ten different gene cargos, consisting of the common therapeutic genes *CEP290*, *HBB*, *PAH*, *GBA*, *ADA*, *SERPINA1* and the *NYESO* (*CTAG*) T cell receptor at the *ACTB* locus and a subset of these cargos at *LMNB1*. These cargos, which varied in size from 969 bp to 4,906 bp, had integration frequencies between 4 and 22% depending on the gene and insertion locus, with minimal indel formation (Fig. 2j and Extended Data Fig. 5e). We next tested if PASTE could insert with base-pair resolution, which is useful for in-frame protein tagging or expressing cargo without disruption of endogenous gene expression. As BxbINT leaves residual sequences in the genome (termed *attL* and *attR*) after cargo integration, we hypothesized that these genomic scars could serve as protein linkers. We positioned the frame of the *attR* sequence through strategic placement of the *attP* on the minicircle cargo, achieving a suitable protein linker, GGLSGQPPRSPSSGSSG. Using this linker, we tagged four genes (*ACTB*, *SRRM2*, *NOLC1* and *LMNB1*)

**Fig. 2 | Evaluating design rules for efficient PASTE insertion at endogenous genomic loci. a**, Schematic of pooled oligonucleotide library design for high-throughput screening of atgRNA designs at endogenous gene targets. **b**, Box plots depicting the editing rates of *attB* addition at the different endogenous targets across 10,580 different atgRNA designs. Boxes indicate between 25th and 75th percentiles, and whiskers indicate 1.5× interquartile range. The center line indicates the 50th percentile. **c**, Scatter plot depicting *attB* site insertion rates versus significance of the editing (−log (*P* value)) as measured by a Student's two-tailed *t*-test against a no Cas9-RT control. **d**, Heat maps depicting percent *attB* site insertion for *LMNB1* guide 1 across different RT, PBS and *attB* lengths. Bar charts indicating normalized summation across relevant PBS, RT or *attB* parameter axes are shown on heat map sides. **e**, Top atgRNA hits from the screen are compared for *attB* site insertion against manually designed atgRNAs (gray bars). **f**, PASTEv3 efficiency for insertion of an EGFP cargo at different endogenous targets is compared between screen-validated atgRNAs and manually designed atgRNAs. **g**, Accuracy results by fivefold cross-validation of an MLP classifier trained on data from the 10,580 atgRNAs. **h**, PASTE integration rates of previously evaluated atgRNAs predicted by the MLP classifier to be efficient (pos. guides) or not efficient (neg. guides). **i**, PASTE integration rates of top atgRNAs predicted to be efficient (dark blue) or not efficient (light blue) by the MLP classifier. The solid line indicates median, and the dotted lines indicate 25th and 75th percentiles. **j**, PASTE integration rates and indel formation for integration of ten therapeutically relevant payloads at the *ACTB* locus. **k**, Endogenous protein tagging with GFP via PASTE by in-frame endogenous gene tagging at four loci (*ACTB*, *SRRM2*, *NOLC1* and *LMNB1*). Immunofluorescence images of representative cells are shown. Cells have nuclear DAPI staining and antibody staining of the labeled proteins (IF stain) to show correlation to the endogenous PASTE tagging signal. Data are shown as mean ± s.e.m.; $n = 3$.

with GFP using PASTEv1. To assess correct gene tagging, we compared the subcellular location of GFP with the tagged gene product by immunofluorescence. For all four targeted loci, GFP colocalized with the tagged gene product as expected, indicating successful tagging (Fig. 2k).

## PASTE efficiencies exceed DSB-based insertion methods

To benchmark PASTE against other gene integration methods, we compared PASTEv3 to DSB-dependent gene integration using either NHEJ (that is, HITI) or HDR[6,7] pathways (Fig. 3a). PASTE had better gene insertion efficiencies than HITI (Fig. 3b). As DSB generation can lead to insertions



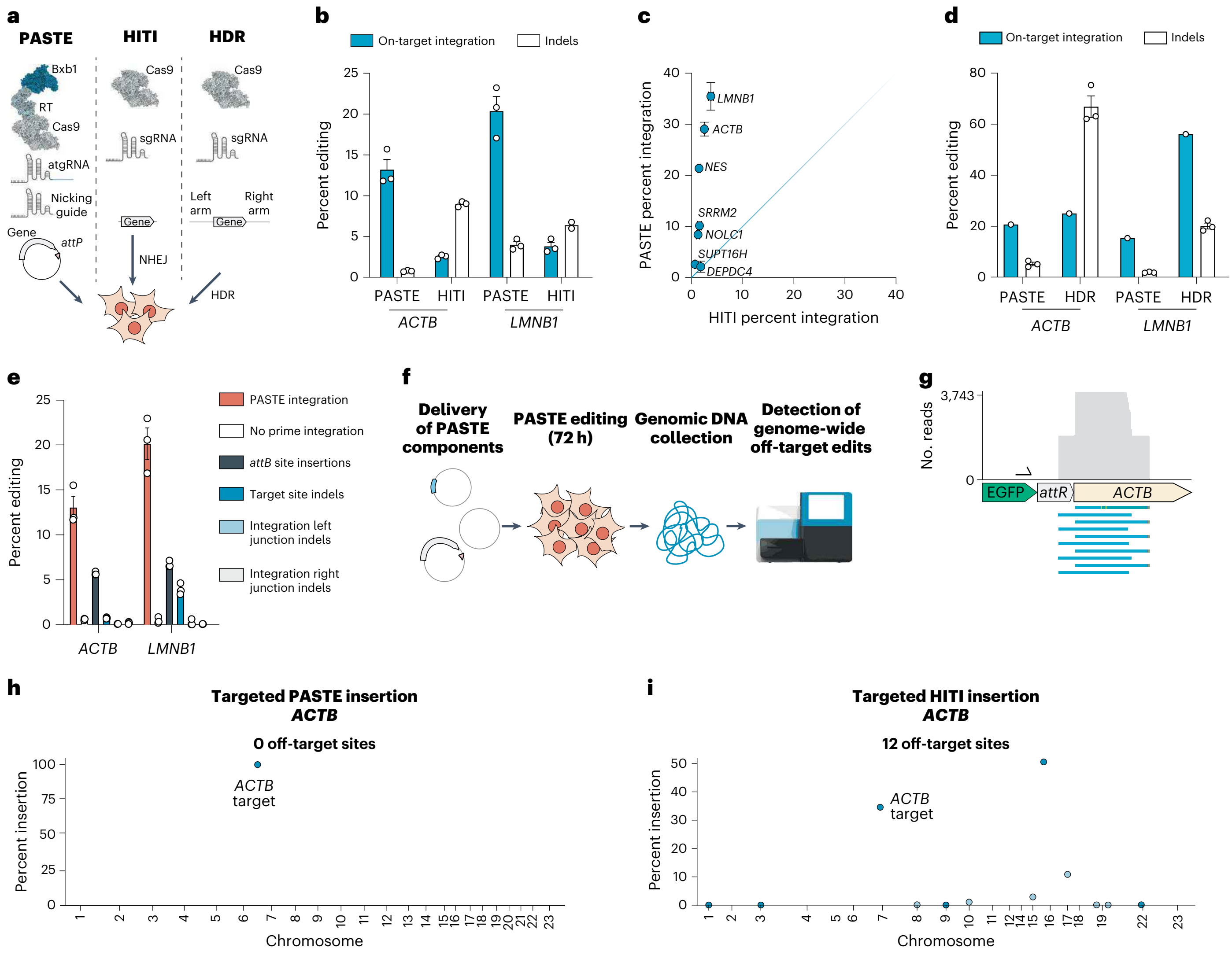


**Fig. 3 | Characterization of genome-wide PASTE specificity and purity of integration compared to other integration approaches.** **a**, Schematic of PASTE, HITI and HDR gene integration approaches; sgRNA, single guide RNA. **b**, Integration of a GFP template by PASTE at the *ACTB* and *LMNB1* loci compared to HITI at the same target. Quantification was performed by ddPCR. Integration efficiency was compared to the rate of byproduct indel generation. **c**, GFP integration efficiency at a panel of genomic loci by PASTE compared to insertion rates with HITI. **d**, Integration of a GFP template by PASTE at the *ACTB* and *LMNB1* loci compared to HDR at the same target. Quantification was performed by single-cell clone counting because HDR homology precludes the use of ddPCR. Integration efficiency was compared to the rate of byproduct indel generation. **e**, Analysis of all possible editing outcomes for PASTEv3 at the *ACTB* and *LMNB1* sites. **f**, Schematic of NGS method to assay genome-wide off-target integration sites by PASTE and HITI. **g**, Alignment of reads at the on-target *ACTB* site using our unbiased genome-wide integration assay, showing expected on-target PASTE integration outcomes. **h**, Manhattan plot of averaged integration events for multiple single-cell clones with PASTE editing. The on-target site is at the *ACTB* gene on chromosome 7 (labeled). The number of off-targets with greater than 0.1% integration is shown. **i**, Manhattan plot of averaged integration events for multiple single-cell clones with HITI editing. The on-target site is at the *ACTB* gene on chromosome 7 (labeled). The number of off-targets with greater than 0.1% integration is shown. Data are shown as mean ± s.e.m.; *n* = 3.

or deletions (indels) as an alternative and undesired editing outcome, we assessed the indel frequency by next-generation sequencing (NGS) and found significantly fewer indels generated with PASTEv3 than HITI in both HEK293FT and HepG2 cells (Fig. 3b and Extended Data Fig. 6a), showcasing the high purity of gene integration outcomes with PASTE due to the lack of DSB formation. On a panel of seven different endogenous targets, PASTEv3 exceeded HITI editing at six of seven genes, with similar efficiency for the seventh gene (Fig. 3c). We also compared PASTEv3 to previously validated HDR constructs at the N terminus of *ACTB* and *LMNB1* for EGFP tagging[42] and found that although PASTE had similar efficiency at the *ACTB* locus and lower efficiency at the *LMNB1* locus, it generated significantly fewer indels than HDR (Fig. 3d). Notably, both HDR and HITI generate more indels than desired on-target integrations at the *ACTB* locus. To comprehensively profile PASTE outcomes, we analyzed all possible intermediate or alternative editing outcomes for PASTEv2 and PASTEv3 at the *ACTB* and *LMNB1* loci, including presence of residual *attB* sites and indels at either end of the integration junction. Residual *attB* sites were a minority event, with an integration frequency into available *attB* sites at ~70–75% (Fig. 3e), and testing the effect of these residual *attB* sites via western blotting showed that they had no effect on protein expression (Extended Data Fig. 6b). Additionally, we found no indel formation at the integration junctions (Fig. 3e).

## Off-target characterization of PASTE and HITI integration

As off-target editing is a critical consideration for genome-editing technologies, we explored the specificity of PASTE at specific sites through two hypotheses: (1) off-targets generated by BxbINT integration into pseudo-*attB* sites in the human genome and (2) off-targets generated via guide- and Cas9-dependent editing in the human genome. While BxbINT lacks documented integration into the human genome at pseudo-attachment sites[43], we computationally identified potential sites with partial similarity to the natural BxbINT *attB* core sequence. We tested for BxbINT integration by digital droplet PCR (ddPCR) across these sites and found no off-target activity (Extended Data Fig. 6c–g). To assay Cas9 off-targets for our *ACTB* atgRNA, we identified two potential off-target sites via computational prediction and found no off-target integration for PASTE (Extended Data Fig. 6c) but substantial off-target activity by HITI at one of the sites (Extended Data Fig. 6d). While PASTE is shown to be specific for our targets, Cas9-based off-target analysis should be performed for each new PASTE target to ensure specificity.

As computationally predicted sites may not account for all possible off-targets, we additionally evaluated genome-wide off-targets due to either Cas9 or BxbINT through tagging and PCR amplification of insert–genomic junctions (Fig. 3f). We isolated single-cell clones for conditions with PASTEv3 integration and negative controls missing PE2, and deep sequencing of insert–genomic junctions from these clones showed all reads aligning to the on-target *ACTB* site, confirming no off-target genomic insertions (Fig. 3g–i and Extended Data Fig. 6h). We also used this genome-wide pipeline to analyze HITI off-targets using the same *ACTB* guide and HITI EGFP insertion template and found substantial off-target activity across the genome, with 12 different sites identified across 10 chromosomes. Moreover, the on-target *ACTB* edit was only 34.8% of the reads identified, with two other off-targets having higher efficiency. These results show that linear template-based integration approaches have significant off-target activity and highlight the benefits of using circular templates with a dual-nicking PASTE system.

Expression of reverse transcriptases and integrases involved in PASTE may have detrimental effects on cellular health. To determine the extent of these effects, we transfected the complete PASTEv1 system, the corresponding guides and cargo with only PE2 and the corresponding guides and cargo with only BxbINT and compared them to both GFP control transfections and guides without protein expression via transcriptome-wide RNA sequencing. We found that, while BxbINT expression in the absence of prime editing had several significant off-targets, the complete PASTE system had only one differentially regulated gene with more than a 1.5-fold change (Extended Data Fig. 6i,j). Genes upregulated by BxbINT overexpression included stress response genes, such as *TENT5C* and *DDIT3*, but these changes were not seen in the expression of the PASTE system (Extended Data Fig. 6i,j), potentially due to differences in expression when BxbINT is linked to the PASTE construct.

### *attB* engineering, gene replacement and multiplexed PASTE

To optimize PASTE efficiency, we profiled attachment site mutants for optimization of integration kinetics of BxbINT, especially for shorter *attB* sites that have reduced integration efficiency. Testing a panel of different *attP* sequences previously shown to affect BxbINT integration[44,45], we found *attP* sequence variants that substantially improved the integration rate (Extended Data Fig. 7a,b). To further improve integration, we expanded our *attP* mutagenesis using a pooled screen to evaluate over 5,775 *attP* variants containing single and double mutations for enhanced integration activity (Fig. 4a and Extended Data Fig. 7c) and found a mutant with improved integration activity over the wild-type (WT) *attP* at both the *ACTB* and *LMNB1* target sites with PASTEv3 (Fig. 4b).

Using the optimized *attP* site mutant 1, we tested whether it might be possible to replace a target sequence by combining integration of a transgene with simultaneous deletion, building on recent developments using prime editing for replacing genomic DNA with short sequences[46,47]. Using paired atgRNAs at the *LMNB1* locus with a 38-bp *attB* sequence and RTs that bridge to the other landing site, we replaced 130 bp and 385 bp of genomic sequence at a rate of 7–10% (Fig. 4c). Combining deletion with integration, we could insert the EGFP payload at ~8% integration efficiency to replace 130 bp of genomic sequence, with higher efficiencies for the *attP* mutant 1 insertion template than the WT *attP* (Fig. 4d). This version of PASTE, which we termed PASTE-Replace, only requires two atgRNAs containing the PBS and *attB* sequences, with an optional inclusion of RT to bridge the deletion. We further profiled PASTE-Replace at the *ACTB*, *NOLC1* and *CCR5* loci, finding that PASTEv3 could insert an EGFP payload at 21%, 25% and 4.5% efficiency, respectively, using both single atgRNA/nicking guide combinations and dual guide RNAs (Extended Data Fig. 7d–f). We also compared PASTE-Replace to the recently published paired guide integrase approach, Twin-PE mediated knock-in[17], finding that PASTE-Replace had 8.3-, 42.1- and 9-fold higher integration efficiencies at *ACTB*, *NOLC1* and *CCR5*, respectively (Extended Data Fig. 7d–f). Quantifying residual *attB* placement, we found that improved efficiency of PASTE-Replace was not primarily driven by the efficiency of *attB* integration, which did not have significant differences between the two approaches (Extended Data Fig. 7e–g). Integrating performance across all three loci, the improved integration efficiency of PASTE-Replace was driven by a combination of the PASTEv3 construct and longer Bxb1 *attB* and *attP* lengths, which are more optimal for integrase efficiency (Extended Data Fig. 7d–f).

The central dinucleotide of the *attP* and *attB* sites of BxbINT is intimately involved in the association of these attachment sites for integration[41], and changing the matched central dinucleotide sequences can modify integrase activity and provide orthogonality for insertion of two genes[48]. We hypothesized that expanding the set of *attB*/*attP* dinucleotides could enable multiplexed gene insertion with PASTE using orthogonal atgRNA combinations (Fig. 4e). To find optimal *attB*/*attP* dinucleotides for PASTE insertion, we profiled the efficiency of GFP integration at the *ACTB* locus with PASTE across all 16 dinucleotide *attB*/*attP* sequence pairs. We found several dinucleotides with integration efficiencies greater than the WT GT sequence (Extended Data Fig. 7h). The majority of dinucleotides had 75% integration efficiency or greater compared to WT *attB*/*attP* efficiency, implying that these dinucleotides could be potential orthogonal channels for multiplexed gene insertion with PASTE.

Next, we explored the specificity of matched and unmatched *attB/attP* dinucleotide interactions. We comprehensively profiled the interactions between all dinucleotide combinations in a scalable fashion using a pooled assay to compare *attB/attP* integration (Extended Data Fig. 7i). By barcoding 16 *attP* dinucleotide plasmids with unique identifiers, cotransfecting this AttP pool with the BxbINT integrase expression vector and a single *attB* dinucleotide acceptor plasmid and sequencing the resulting integration products, we measured the relative integration efficiencies of all possible *attB/attP* pairs (Extended Data Fig. 7j). We found that dinucleotide specificity varied wildly, with some dinucleotides (GG) exhibiting strong self-interaction with negligible cross-talk and others (AA) showing minimal self-preference. Sequence logos of *attP* preferences (Extended Data Fig. 7k) revealed that dinucleotides with C or G in the first position have stronger preferences for *attB* dinucleotide sequences with shared first bases, while other *attP* dinucleotides, especially those with an A in the first position, have reduced specificity for the first *attB* base.

Informed by the efficiency and specificity of the central dinucleotides, we tested GA, AG, AC and CT dinucleotide atgRNAs for GFP integration at *ACTB* with PASTEv3, either paired with their corresponding *attP* cargo or mispaired with the other three dinucleotide *attP* sequences. We found that all four of the tested dinucleotides efficiently integrated cargo only when paired with the corresponding *attB/attP* pair, with no detectable integration across mispaired combinations (Fig. 4f).

Selecting the three top dinucleotide attachment site pairs (CT, AG and GA), we designed atgRNAs that target *ACTB* (CT), *LMNB1* (AG) and *NOLC1* (GA) and corresponding minicircle cargo containing GFP (CT), mCherry (AG) and yellow fluorescent protein (YFP; GA). After co-delivering these reagents to cells, we found that we could achieve

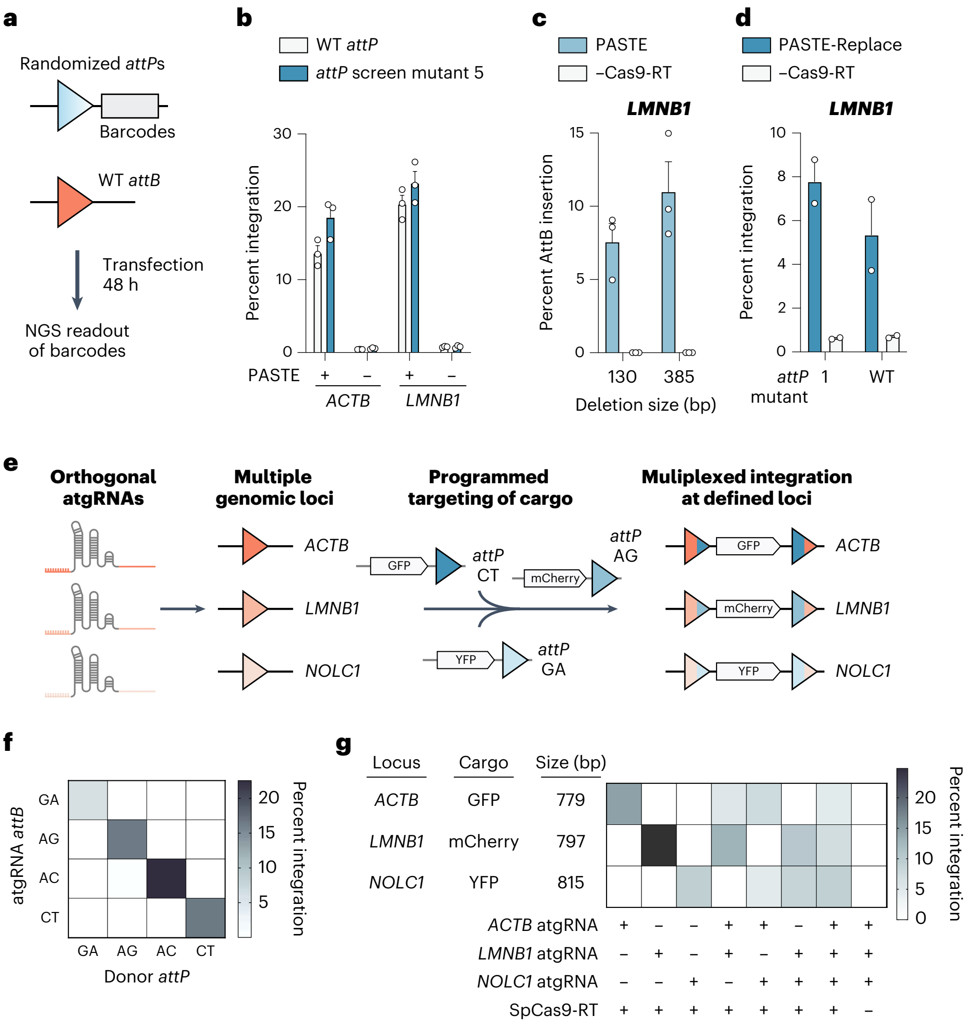


**Fig. 4 | Multiplexed and orthogonal gene insertion with PASTE. a**, Schematic for *attP* mutagenesis screen for identifying *attP* mutants that promote higher integration efficiencies with PASTE. **b**, Evaluation of two *attP* variants from the pooled screen for PASTE integration activity at the *ACTB* and *LMNB1* loci. **c**, *attB* site replacement efficiency with the PASTE-Replace system at the *LMNB1* locus. **d**, EGFP gene replacement efficiency with the PASTE-Replace system at the *LMNB1* locus using payloads with either *attP* mutant 1 or WT *attP*. **e**, Schematic of multiplexed integration of different cargo sets at specific genomic loci. Three fluorescent cargos (GFP, mCherry and YFP) are inserted orthogonally at three different loci (*ACTB*, *LMNB1* and *NOLC1*) for in-frame gene tagging. **f**, Orthogonality of the top four *attB*/*attP* dinucleotide pairs evaluated for GFP integration with PASTE at the *ACTB* locus. **g**, Efficiency of multiplexed PASTE insertion of combinations of fluorophores at *ACTB*, *LMNB1* and *NOLC1* loci. Data are shown as mean ± s.e.m.; $n = 3$.

single-plex, dual-plex and tri-plex editing, as read out by bulk genomic DNA collection and ddPCR, of all possible combinations of these atgRNAs and cargo in the range of 5–25% integration with PASTEv1 (Fig. 4g).

A useful application for multiplexed gene integration is for labeling different proteins to visualize intracellular localization and interactions within the same cell. We used PASTEv1 to simultaneously tag *ACTB* (GFP) and *NOLC1* (mCherry) or *ACTB* (GFP) and *LMNB1* (mCherry) in the same cell. We observed that no overlap of GFP and mCherry fluorescence existed, and we confirmed that tagged genes were visible in their appropriate cellular compartments based on the known subcellular localizations of the *ACTB*, *NOLC1* and *LMNB1* protein products (Extended Data Fig. 7l).

Programmable gene integration provides a modality for expression of therapeutic protein products, and we tested protein production of therapeutically relevant proteins alpha-1 antitrypsin (encoded by *SERPINA1*) and carbamoyl-phosphate synthetase I (encoded by *CPS1*), involved in the diseases alpha-1 antitrypsin deficiency and CPS1 deficiency, respectively. By tagging gene products with the luminescent protein subunit HiBiT[49], we could independently assess transgene production and secretion in response to PASTE treatment (Extended Data Fig. 8a). We transfected PASTEv1 with *SERPINA1* or *CPS1* cargo in HEK293FT cells and a human hepatocellular carcinoma cell line (HepG2) and found efficient integration at the *ACTB* locus (Extended Data Fig. 8b,c). This integration resulted in robust protein expression, intracellular accumulation of transgene products and secretion of proteins into the medium (Extended Data Fig. 8d–g).

## Discovery and development of serine integrases for PASTE

As we found that integrase choice can have implications for integration activity (Fig. 1c,d), we decided to mine bacterial and metagenomic sequences for new phage-associated serine integrases (Fig. 5a). Exploring over 10 terabytes worth of data from National Center for Biotechnology Information (NCBI), Joint Genome Institute (JGI) and other sources, we found 25,614 integrases containing the putative catalytic residues (Fig. 5b,c and Extended Data Fig. 3) and annotated their associated attachment sites by evaluating the presence of repetitive structures in potential 50-bp attachment sites near phage boundaries. Analysis of the integrase sequences led to the identification of

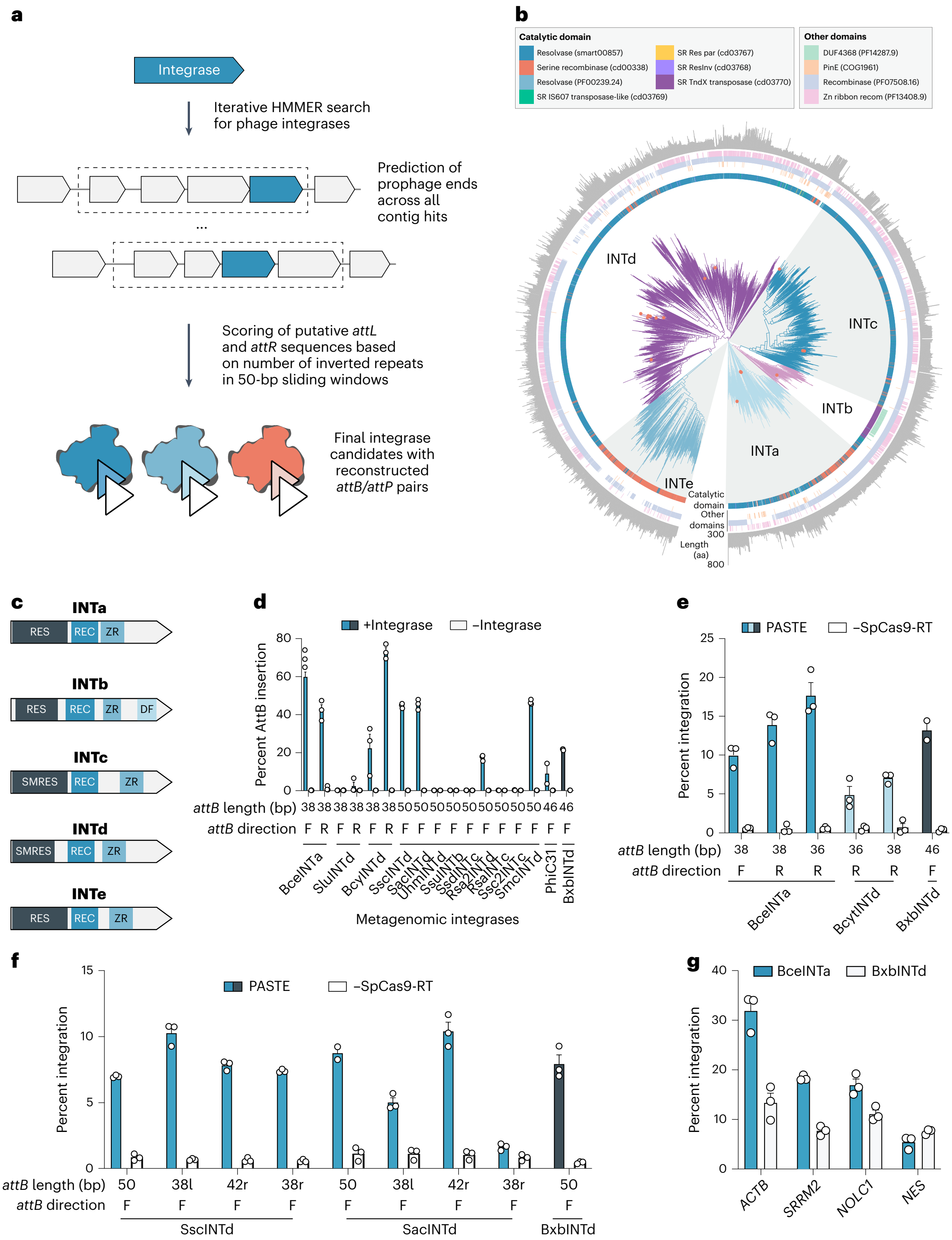


**Fig. 5 | Discovery of phage-derived integrases for programmable gene integration with PASTE. a**, Schematic of integrase discovery pipeline from bacterial and metagenomic sequences. **b**, Phylogenetic tree of discovered integrases showing distinct subfamilies. Synthesized orthologs are shown as orange dots; aa, amino acids. **c**, Domain architecture of the five integrase subfamilies; RES, resolvase (cd00338); REC, recombinase (PF07508); ZR, zinc ribbon (PF13408); DF, unknown domain (DUF4368); SMRES, resolvase (smart00857). **d**, Screening integrase integration activity using reporters in HEK293FT cells compared to BxbINT and phiC31. **e**, PASTE integration activity with BceINT and BcyINT with truncated attachment sites compared to BxbINT. **f**, PASTE integration activity with SscINT and SacINT with truncated attachment sites compared to BxbINT. **g**, Integration of EGFP at different endogenous gene targets for PASTE with either BceINT or BxbINT. Data are shown as mean ± s.e.m.; *n* = 3.

five distinct clusters: INTa–INTe with diverse domain architectures (Fig. 5c). About 20% of the integrases (5,203) derived from metagenomic sequences, presumably from prophages, and 4,452 of these specifically derived from human microbiome metagenomic samples. An initial screen of integrase activity using a reporter system revealed that several integrases were active in HEK293FT cells,

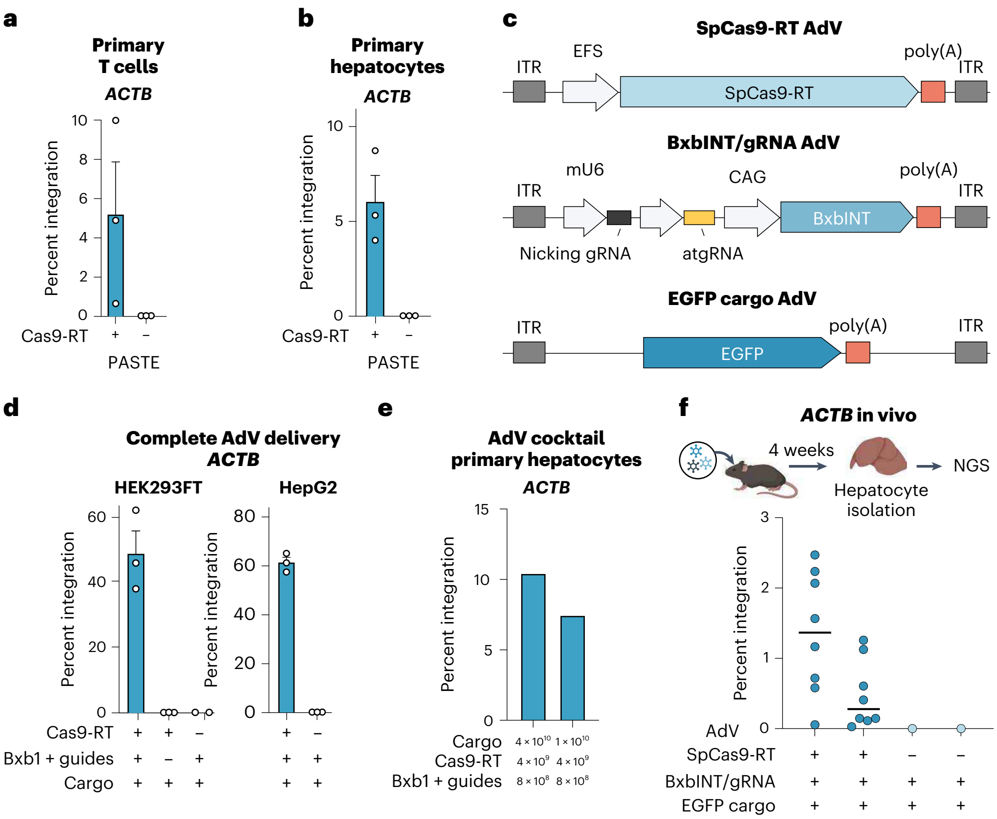


**Fig. 6 | PASTE is compatible with multiple delivery approaches and can be delivered to primary cell types and in vivo animal models. a**, PASTE integration efficiency with single-vector designs in primary human T cells. Data are shown as mean ± s.e.m.; *n* = 3. **b,** PASTE integration efficiency with single-vector designs in primary human hepatocytes. Data are shown as mean ± s.e.m.; *n* = 3. **c**, Schematic of the AdV constructs used to deliver PASTE and the EGFP payload template; ITR, inverted terminal repeat. **d**, AdV delivery of all PASTE components in HEK293FT and HepG2 cells. Data are shown as mean ± s.e.m.; *n* = 3. **e**, Integration efficiency of AdV delivery of integrase, guides and cargo in primary human hepatocytes (PXB-cells). Viral components were listed at the dosages indicated; *n* = 1. **f**, AdV EGFP template integration efficiency at the human *ACTB* locus in the liver of a liver-humanized mouse model using adenovirally delivered PASTE. Integration efficiency was measured 4 weeks after injection. For integration conditions, points represent different regions of the liver analyzed for editing. A schematic for in vivo targeted gene integration with PASTE via retroorbital injection is shown at the top. Data are shown as mean (*n* = 8).

including multiple with more activity than BxbINT, a member of the INTd family (Fig. 5d). Using the predicted 50-bp sequences encoded in atgRNAs along with minicircles containing the complementary *attP* sites, we found that these integrases were compatible with PASTE but performed less effectively than BxbINTd-based PASTE (Extended Data Fig. 8h). We hypothesized that this reduction in performance of the new integrases was due to their longer 50-bp *attB* sequences, and so we explored truncations of these *attB*s in the hopes of finding more minimal attachment sites. Truncation screening on integrase reporters revealed that *attB* truncations of all the integrases, including as short as 34 bp, were still active, and many had more activity than BxbINTd (Extended Data Fig. 8i). After porting these new shorter a*ttB*s to atgRNAs for PASTE, we found that several integrases had more activity in the PASTE system than BxbINTd-based PASTE at the *ACTB* locus, including the integrase from *Bacillus cereus* (BceINTa), an integrase from a stool sample from China (SscINTd) and an integrase from a stool sample from an adult in China (SacINTd), while others, like the integrases from *B. cytotoxicus* (BcyINTd) and *Staphylococcus lugdunensis* (SluINTd), did not (Fig. 5e,f). Additionally, we computationally nominated a set of integrases with shorter *attB* sites of 30 nucleotides and tested them as PASTE and found that several candidates, Sss2INTd and SscINTd, functioned as a complete PASTE system. To improve PASTE with our new integrases, we fused BceINTa to SpCas9-MLV-RT$^{L139P}$, termed PASTEv4, and found that it performed better than BxbINTd-based PASTE across several endogenous gene loci (Fig. 5g).

## PASTE efficiency in non-dividing and primary cells

As PASTE does not rely on DSB repair pathways that are only active in dividing cells, we tested PASTE activity in non-dividing cells by transfecting either Cas9 and HDR templates or PASTE into HEK293FT cells and arresting cell division[50] via aphidicolin treatment (Extended Data Fig. 9a). In this model of blocked cell division, we found that PASTEv1 maintained GFP gene integration activity greater than 20% at the *ACTB* locus, whereas HDR-mediated integration was abolished (Extended Data Fig. 9b,c). To evaluate the size limits for therapeutic transgenes, we evaluated insertion of cargos up to 13.3 kb in length in both dividing and aphidicolin-treated cells and found insertion efficiency greater than 10% (Extended Data Fig. 9d). To overcome reduction of large insert delivery to cells due to potential delivery inefficiencies, we found that delivering larger DNA insert amounts could significantly improve gene integration efficiency (Extended Data Fig. 9e).

We also expanded PASTE editing to additional cell types and tested PASTE in the K562 lymphoblast line, primary human T cells and primary human hepatocytes. We found that PASTEv1 had ~15% gene integration activity in K562 cells and around 5% efficiency in primary human T cells (Fig. 6a and Extended Data Fig. 9f). In addition, in non-dividing quiescent human primary hepatocytes, we found that PASTEv1 was capable of ~5% gene integration at the *ACTB* locus (Fig. 6b) after sorting for transfected cells, consistent with the non-dividing activity we observed with the aphidicolin-treated HEK293FT cells.

### Viral therapeutic payload delivery with PASTE

To explore compatibility of PASTEv3 with therapeutically relevant delivery modalities, we explored whether components of the PASTE system could be delivered with either AAV or AdV vectors. Testing AAV-delivered cargo with an *attP*-containing payload in conjunction with other PASTE components delivered via transfection, we found ~4–10% integration of the viral payload in a dose-dependent fashion (Extended Data Fig. 9g–i). The AAV genome serving as a suitable template for serine integrase-mediated insertion is consistent with reports of AAV genome circularization in cells[51].

To package larger cargos in viral vectors, we used an AdV vector, an emerging approach for clinical delivery of large genes[52]. We evaluated whether AdV could deliver a suitable template for BxbINT-mediated insertion along with plasmids for PASTEv3 and guide expression or AdV delivery of guides and BxbINT with plasmid delivery of SpCas9-RT. We found that we could achieve 10–20% integration of the ~36-kb AdV genome carrying EGFP in HEK293FT and HepG2 cells (Extended Data Fig. 9j).

To further demonstrate that PASTE would be amenable for in vivo delivery, we developed an mRNA version of the PASTEv1 protein components and chemically modified synthetic atgRNA and nicking guide against the *LMNB1* target. Electroporation of the mRNA and guides along with delivery of the template via AdV or plasmid yielded high-efficiency integration up to ~20% (Extended Data Fig. 9k–m). As we hypothesized that more sustained BxbINT expression would allow for integration into newly placed *attB* sites in the genome, we tested circular mRNA expression[53] and found that this boosted the efficiency of integration to ~30% (Extended Data Fig. 10a).

To package the complete PASTE system in viral vectors, we devised a vector strategy to package the PASTEv1 components across two additional AdV vectors, allowing the cargo and PASTEv1 system to be delivered across three AdV vectors (Fig. 6c). We found that the complete PASTE system (Cas9-RT, integrase and guide RNAs and cargo) could be substituted by AdV delivery, with integration of up to ~50–60% with viral-only delivery in HEK293FT and HepG2 cells (Fig. 6d and Extended Data Fig. 10b). As an evaluation of therapeutic feasibility of adenovirally delivered PASTE, we tested complete AdV delivery at three different cargo amounts in primary human hepatocytes (PXB-cells) and found editing efficiencies up to 10.5% in a cargo-dependent fashion using an NGS-based integration analysis, with up to 3.8% integration using an AAV template (Fig. 6e and Extended Data Fig. 10c,d).

### In vivo delivery of PASTE for liver gene integration

We next applied AdV PASTE delivery for in vivo targeting of the liver. As our AdV PASTE components were designed to target the human *ACTB* locus, we performed experiments in ~5.5-month-old, liver-humanized FRG mice (*Fah*$^{-/-}$*Rag2*$^{-/-}$*Il2rg*$^{-/-}$ on C587BL/6 with ≥70% human hepatocyte repopulation[54]). Mice were retro-orbitally injected with the triple-vector PASTE cocktail and maintained for 3 weeks before liver collection and NGS-based integration analysis. We found that PASTE was capable of integration rates as high as 2.5% in the human hepatocytes in the chimeric liver, with indel byproduct formation at the *ACTB* locus of 0.1–0.2% (Fig. 6f and Extended Data Fig. 10e–i).

## Discussion

We developed PASTE by engineering of Cas9, reverse transcriptase and integrase linkers to create a fusion protein capable of efficient integration (5–50%) of diverse cargos at precisely defined target locations within the human genome with small, stereotyped scars that can serve as protein linkers. We demonstrate the versatility of PASTE for gene tagging, gene replacement, gene delivery and protein production and secretion. Through extensive characterization of integrase attachment sites, we engineered multiplexed gene integration with PASTE, enabling applications such as the specific fusion of three different endogenous genes with three different fluorescent cargos. Overall, we show PASTE insertions at 9 different endogenous sites with 13 different cargos ranging in size from 779 bp to ~36,000 bp, which would enable insertion of greater than 99.7% of human cDNAs as transgenes[55]. We additionally benchmarked PASTE against other prime-editing and integrase-based insertion approaches[17] and found significant improvements driven by a combination of more optimized *attB* and *attP* sequences and the fusion-based design of the PASTEv3 editor. In agreement with previous studies of serine integrases and prime editing, we found no off-target activity with PASTE.

Metagenomic mining enabled the discovery of thousands of putative integrase/attachment site combinations and engineering of multiple integrase orthologs with improved activity and reduced attachment site requirements to further optimize the activity of PASTE, generating a PASTEv4 system using the BceINT integrase. We anticipate that the compendium of 25,614 serine recombinases that we discovered and characterized will be useful for additional PASTE and synthetic biology applications, although more work is needed to fully characterize the activity of these integrases and any natural pseudosites of integration in the human genome that might serve as off-target sites. Moreover, in contrast to transposase-based integration systems[56,57], PASTE integration is stereotyped, allowing for precise design of integration and predictable gene fusions. As PASTE does not rely on HDR, it can function in non-dividing cells, including in primary hepatocytes and T cells, and we demonstrate human hepatocyte editing in vivo via AdV delivery to liver-humanized mice. In addition, as delivery conditions were not optimized, and AdV can be hepatotoxic, we anticipate that in vivo activity can be substantially improved.

Programmable insertion is a fundamental tool for genetics for applications such as tagging of gene products, interrogating variants of unknown function and developing disease models. PASTE also enables therapeutic correction of genetic disease through insertion of full-length, functional genes at native loci, a viable strategy for both treating recessive loss of function mutations that cover 4,122 genetic diseases[58] and overcoming dominant-negative mutations. Current genome-editing approaches for diseases such as cystic fibrosis or Leber's congenital amaurosis[59] are limited, as systems must be tailored for specific mutations[60,61], requiring unique genome-editing therapies for each subset of the clinical population. Programmable insertion of the WT gene at the endogenous location could address most potential mutations, serving as a blanket therapy. Beyond direct correction of hereditary disease, gene insertion provides a promising avenue for cell therapies, and efficient integration of engineered transgenes, such as chimeric antigen receptors at specific loci, can produce improved therapeutic products in comparison to random integration[62].

The development of PASTE marries engineering of CRISPR nucleases with the discovery and mammalian characterization of a variety of serine integrases with diverse sequence preferences. By providing efficient, multiplexed integration of transgenes in dividing and non-dividing cells and in animal models, the PASTE platform builds on fundamental developments in both integrase and CRISPR biology to expand the scope of genome editing and enable new applications across basic biology and therapeutics.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01527-4.

## References

1. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR–Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).

2. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824–844 (2020).
3. Wright, A. V., Nuñez, J. K. & Doudna, J. A. Biology and applications of CRISPR systems: harnessing nature's toolbox for genome engineering. *Cell* **164**, 29–44 (2016).
4. Nami, F. et al. Strategies for in vivo genome editing in nondividing cells. *Trends Biotechnol.* **36**, 770–786 (2018).
5. Suzuki, K. et al. In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. *Nature* **540**, 144–149 (2016).
6. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
7. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
8. Rouet, P., Smih, F. & Jasin, M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol. Cell. Biol.* **14**, 8096–8106 (1994).
9. Chapman, J. R., Taylor, M. R. G. & Boulton, S. J. Playing the end game: DNA double-strand break repair pathway choice. *Mol. Cell* **47**, 497–510 (2012).
10. Geisinger, J. M. & Stearns, T. CRISPR/Cas9 treatment causes extended TP53-dependent cell cycle arrest in human cells. *Nucleic Acids Res.* **48**, 9067–9081 (2020).
11. Wang, H. et al. Development of a self-restricting CRISPR–Cas9 system to reduce off-target effects. *Mol. Ther. Methods Clin. Dev.* **18**, 390–401 (2020).
12. Kanca, O. et al. An efficient CRISPR-based strategy to insert small and large fragments of DNA using short homology arms. *eLife* **8**, e51539 (2019).
13. Gaudelli, N. M. et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
14. Rees, H. A. & Liu, D. R. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* **19**, 770–788 (2018).
15. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
16. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
17. Anzalone, A. V. et al. Programmable deletion, replacement, integration and inversion of large DNA sequences with twin prime editing. *Nat. Biotechnol.* **40**, 731–740 (2021).
18. Wang, J. et al. Efficient targeted insertion of large DNA fragments without DNA donors. *Nat. Methods* **19**, 331–340 (2022).
19. Ivics, Z., Hackett, P. B., Plasterk, R. H. & Izsvák, Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501–510 (1997).
20. Brown, W. R. A., Lee, N. C. O., Xu, Z. & Smith, M. C. M. Serine recombinases as tools for genome engineering. *Methods* **53**, 372–379 (2011).
21. Calos, M. P. The C31 integrase system for gene therapy. *Curr. Gene Ther.* **6**, 633–645 (2006).
22. Mulholland, C. B. et al. A modular open platform for systematic functional studies under physiological conditions. *Nucleic Acids Res.* **43**, e112 (2015).
23. Ehrhardt, A., Engler, J. A., Xu, H., Cherry, A. M. & Kay, M. A. Molecular analysis of chromosomal rearrangements in mammalian cells after øC31-mediated integration. *Hum. Gene Ther.* **17**, 1077–1094 (2006).
24. Liu, J., Jeppesen, I., Nielsen, K. & Jensen, T. G. Phi c31 integrase induces chromosomal aberrations in primary human fibroblasts. *Gene Ther.* **13**, 1188–1190 (2006).
25. Kovač, A. et al. RNA-guided retargeting of Sleeping Beauty transposition in human cells. *eLife* **9**, e53868 (2020).
26. Ma, S. et al. Enhancing site-specific DNA integration by a Cas9 nuclease fused with a DNA donor-binding domain. *Nucleic Acids Res.* **48**, 10590–10601 (2020).
27. Chen, S. P. & Wang, H. H. An engineered Cas–transposon system for programmable and site-directed DNA transpositions. *CRISPR J.* **2**, 376–394 (2019).
28. Bhatt, S. & Chalmers, R. Targeted DNA transposition in vitro using a dCas9–transposase fusion protein. *Nucleic Acids Res.* **47**, 8126–8135 (2019).
29. Hew, B. E., Sato, R., Mauro, D., Stoytchev, I. & Owens, J. B. RNA-guided piggyBac transposition in human cells. *Synth. Biol.* **4**, ysz018 (2019).
30. Chaikind, B., Bessen, J. L., Thompson, D. B., Hu, J. H. & Liu, D. R. A programmable Cas9–serine recombinase fusion protein that operates on DNA sequences in mammalian cells. *Nucleic Acids Res.* **44**, 9758–9770 (2016).
31. Akopian, A., He, J., Boocock, M. R. & Stark, W. M. Chimeric recombinases with designed DNA sequence recognition. *Proc. Natl Acad. Sci. USA* **100**, 8688–8691 (2003).
32. Gordley, R. M., Smith, J. D., Gräslund, T. & Barbas, C. F. III Evolution of programmable zinc finger-recombinases with activity in human cells. *J. Mol. Biol.* **367**, 802–813 (2007).
33. Mercer, A. C., Gaj, T., Fuller, R. P. & Barbas, C. F. III Chimeric TALE recombinases with programmable DNA sequence specificity. *Nucleic Acids Res.* **40**, 11163–11172 (2012).
34. Gersbach, C. A., Gaj, T., Gordley, R. M., Mercer, A. C. & Barbas, C. F. III Targeted plasmid integration into the human genome by an engineered zinc-finger recombinase. *Nucleic Acids Res.* **39**, 7868–7878 (2011).
35. Prorocic, M. M. et al. Zinc-finger recombinase activities in vitro. *Nucleic Acids Res.* **39**, 9316–9328 (2011).
36. Gordley, R. M., Gersbach, C. A. & Barbas, C. F. III Synthesis of programmable integrases. *Proc. Natl Acad. Sci. USA* **106**, 5053–5058 (2009).
37. Xu, Z. et al. Accuracy and efficiency define Bxb1 integrase as the best of fifteen candidate serine recombinases for the integration of DNA into the human genome. *BMC Biotechnol.* **13**, 87 (2013).
38. Kay, M. A., He, C. -Y. & Chen, Z. -Y. A robust system for production of minicircle DNA vectors. *Nat. Biotechnol.* **28**, 1287–1289 (2010).
39. Dang, Y. et al. Optimizing sgRNA structure to improve CRISPR–Cas9 knockout efficiency. *Genome Biol.* **16**, 280 (2015).
40. Oscorbin, I. P., Wong, P. F. & Boyarskikh, U. A. The attachment of a DNA-binding Sso7d-like protein improves processivity and resistance to inhibitors of M-MuLV reverse transcriptase. *FEBS Lett.* **594**, 4338–4356 (2020).
41. Ghosh, P., Kim, A. I. & Hatfull, G. F. The orientation of mycobacteriophage Bxb1 integration is solely dependent on the central dinucleotide of *attP* and *attB*. *Mol. Cell* **12**, 1101–1111 (2003).
42. Sun, D. et al. A functional genetic toolbox for human tissue-derived organoids. *eLife* **10**, e67886 (2021).
43. Keravala, A. et al. A diversity of serine phage integrases mediate site-specific recombination in mammalian cells. *Mol. Genet. Genomics* **276**, 135–146 (2006).
44. Singh, S., Ghosh, P. & Hatfull, G. F. Attachment site selection and identity in Bxb1 serine integrase-mediated site-specific recombination. *PLoS Genet.* **9**, e1003490 (2013).
45. Zhang, Q., Azarin, S. M. & Sarkar, C. A. Model-guided engineering of DNA sequences with predictable site-specific recombination rates. *Nat. Commun.* **13**, 4152 (2022).
46. Jiang, T., Zhang, X. -O., Weng, Z. & Xue, W. Deletion and replacement of long genomic sequences using prime editing. *Nat. Biotechnol.* **40**, 227–234 (2021).
47. Choi, J. et al. Precise genomic deletions using paired prime editing. *Nat. Biotechnol.* **40**, 218–226 (2022).

48. Jusiak, B. et al. Comparison of integrases identifies Bxb1-GA mutant as the most efficient site-specific integrase system in mammalian cells. *ACS Synth. Biol.* **8**, 16–24 (2019).
49. Schwinn, M. K. et al. CRISPR-mediated tagging of endogenous proteins with a luminescent peptide. *ACS Chem. Biol.* **13**, 467–474 (2018).
50. Lin, S., Staahl, B. T., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife* **3**, e04766 (2014).
51. Schnepp, B. C., Jensen, R. L., Chen, C. -L., Johnson, P. R. & Clark, K. R. Characterization of adeno-associated virus genomes isolated from human tissues. *J. Virol.* **79**, 14793–14803 (2005).
52. Wold, W. S. M. & Toth, K. Adenovirus vectors for gene therapy, vaccination and cancer gene therapy. *Curr. Gene Ther.* **13**, 421–433 (2013).
53. Wesselhoeft, R. A., Kowalski, P. S. & Anderson, D. G. Engineering circular RNA for potent and stable translation in eukaryotic cells. *Nat. Commun.* **9**, 2629 (2018).
54. Azuma, H. et al. Robust expansion of human hepatocytes in *Fah*$^{-/-}$/*Rag2*$^{-/-}$/*Il2rg*$^{-/-}$ mice. *Nat. Biotechnol.* **25**, 903–910 (2007).
55. Bateman, A. et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2020).
56. Strecker, J. et al. RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48–53 (2019).
57. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).
58. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).
59. Maeder, M. L. et al. Development of a gene-editing approach to restore vision loss in Leber congenital amaurosis type 10. *Nat. Med.* **25**, 229–233 (2019).
60. Mackay, D. S. et al. Screening of a large cohort of Leber congenital amaurosis and retinitis pigmentosa patients identifies novel *LCA5* mutations and new genotype–phenotype correlations. *Hum. Mutat.* **34**, 1537–1546 (2013).
61. Marson, F. A. L., Bertuzzo, C. S. & Ribeiro, J. D. Classification of *CFTR* mutation classes. *Lancet Respir. Med.* **4**, e36 (2016).
62. Eyquem, J. et al. Targeting a CAR to the *TRAC* locus with CRISPR/Cas9 enhances tumour rejection. *Nature* **543**, 113–117 (2017).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Methods

### Cloning of atgRNAs and nicking guides

atgRNA and nicking guides were cloned by Golden Gate assembly of PCR products. Guide products were amplified by PCR (KAPA HiFi HotStart DNA polymerase, Roche) off of the Cas9 single guide RNA scaffold, with the forward primer containing spacer sequences and the reverse primer containing desired PBS, RT and *attB* insertion sequences, in the case of the atgRNA. PCR products were purified by gel extraction (Monarch gel extraction kit, NEB) and assembled in a Golden Gate assembly containing 6.25 ng of pU6-atgRNA-GG-acceptor (Addgene, 132777), purified PCR product (approximately two- to fourfold molar excess), 0.125 µl of Fermentas Eco31I (Thermo Fisher Scientific), 0.0625 µl of T7 DNA ligase (Enzymatics), 0.0625 µl of 20 mg ml$^{-1}$ bovine serum albumin (NEB), 2× reaction ligation buffer (Enzymatics) and water, for a 6.25-µl total reaction volume. Reactions were incubated between 37 °C and 20 °C for 5 min each for a total of 15 cycles. Two microliters of assembled reactions was transformed into 20 µl of competent Stbl3 cells generated by Mix and Go! competency kit (Zymo) and plated on agar plates supplemented with appropriate antibiotics. After overnight growth at 37 °C, colonies were picked into Terrific Broth (TB) medium (Thermo Fisher Scientific) and incubated with shaking at 37 °C for 24 h. Cultures were collected using a QIAprep Spin Miniprep kit (Qiagen) according to the manufacturer's instructions. All guides used in experiments are summarized in Supplementary Table 4, and all *attB* sequences used in the paper are listed in Supplementary Table 5.

### Cloning of PASTE and cargo constructs

Expression constructs for Cas9-RT fusions, RT mutants, integrases and recombinases and Bxb1 mutants were cloned for mammalian expression via Gibson cloning using Hifi Assembly mix (NEB) according to manufacturer's instructions. All enzyme expression plasmids used in mammalian experiments are summarized in Supplementary Table 6. Sequences of linkers used are listed in Supplementary Table 2, and sequences of Bxb1 and RT mutants are listed in Supplementary Table 3. For cloning of minicircle cargo plasmids, the Bxb1 or equivalent integrase/recombinase *attP* sites and the cargo sequence were introduced into a minicircle parental plasmid with Gibson cloning using Hifi Assembly mix (NEB) according to manufacturer's instructions. The parental plasmid was digested for the sequences to be cloned between the bacterial *attB* and *attP* sites recognized by the ZYCY10P3S2T *Escherichia coli* Minicircle Strain (Systems Bioscience). All transgene and cargo plasmids used in experiments are summarized in Supplementary Table 7.

For all Gibson clonings, 2 µl of assembled reactions was transformed into 20 µl of competent Stbl3 cells generated by Mix and Go! competency kit (Zymo) and plated on agar plates supplemented with appropriate antibiotics. After growth overnight at 37 °C, colonies were picked into TB medium (Thermo Fisher Scientific) and incubated with shaking at 37 °C for 24 h. Cultures were collected using a QIAprep Spin Miniprep kit (Qiagen) according to the manufacturer's instructions.

For screening integrases discovered computationally, gene fragments were synthesized by Twist Biosciences. These genes were then cloned into separate expression vectors for comparing activity of reporters in mammalian cells, and top integrases were cloned into PASTE vectors fused to SpCas9-RT constructs.

### Minicircle production

To produce the minicircle plasmids containing only the integrase *attP* site and the transgene sequence, the parental plasmid was transformed into the ZYCY10P3S2T *E. coli* Minicircle Strain (System Biosciences, MN900A-1) overnight at 37 °C. The next day, a colony was picked into TB medium containing kanamycin antibiotic and grown for approximately 12 h at 32 °C in an incubator shaker. When the optical density at 600 nm reached 4–6, the induction medium was added in a 1:1 ratio to the sample. For the preparation of 100.5 ml of induction medium, 100 ml of lysogeny broth medium (Thermo Fisher Scientific) was mixed with 400 µl of 10 M sodium hydroxide solution (Sigma-Aldrich) and 100 µl of 20% L-arabinose (Sigma-Aldrich). The induced bacterial culture was then incubated at 32 °C in the shaker for 4–5 h. After centrifuging at 5,000*g* for 15 min, the medium was removed, leaving only the cell pellet at the bottom of the tube. For purification of the DNA plasmid, an endotoxin-free plasmid midiprep DNA purification was performed using a NucleoBond Xtra Midi EF kit (Takara Bio) following the manufacturer's protocol. Minicircle digestion was then confirmed using restriction enzymes and subsequent gel electrophoresis that allowed for interpretation of the minicircle and parent plasmid fractions in the purified DNA.

### Mammalian cell culture

HEK293FT cells (Thermo Fisher, R70007) were cultured in Dulbecco's modified Eagle medium with high glucose, sodium pyruvate and GlutaMAX (Thermo Fisher Scientific) and supplemented with 10% (vol/vol) fetal bovine serum (FBS) and 1× penicillin–streptomycin (Thermo Fisher Scientific). For puromycin selection, HEK293FT cells were replated at a 1:3 dilution 1 d after transfection into medium supplemented with 1 µg ml$^{-1}$ final concentration of puromycin (Thermo Fisher Scientific). HEPG2 cells (ATCC, HB8065) were seeded in Eagle's minimum essential medium (Thermo Fisher Scientific) supplemented with 10% (vol/vol) FBS and incubated at 37 °C and 5% $CO_2$. Adherent cells were maintained at a confluency below 80–90% at 37 °C and 5% $CO_2$. K562 cells (ATCC, CCL-243) were cultured in Gibco Roswell Park Memorial Institute 1640 medium (Thermo Fisher Scientific) supplemented with 10% (vol/vol) FBS and maintained at 37 °C and 5% $CO_2$. Primary human peripheral blood CD8$^+$ T cells (Stemcell Technologies, 70027) were expanded using fresh complete ImmunoCult-XF T Cell Expansion medium (Stemcell Technologies, 10981) additionally supplemented with cytokines (human recombinant interleukin-2 (IL-2), Stemcell Technologies, 78036). To stimulate the cells, 25 µl ml$^{-1}$ ImmunoCult human CD3/CD28/CD2 T cell activator (Stemcell Technologies, 10970) was used. Primary human hepatocytes pooled from five donors (Thermo Fisher Scientific, HMCPP5) were plated on 96-well plates coated using collagen I, rat tail (Thermo Fisher Scientific, A10483-01), and transfected 24 h after plating. Stock collagen I was diluted to 50 µg ml$^{-1}$ with 20 mM acetic acid (A6283) and added to plates at 5 µg cm$^{-2}$. Plates were incubated at room temperature for 1 h and rinsed three times with sterile 1× PBS. Thawed hepatocytes were transferred into hepatocyte thaw medium (Thermo Fisher Scientific, CM7500) and centrifuged at 100*g* for 10 min at room temperature. Pelleted cells were resuspended and plated at 2.5 × 10$^4$ using William's E medium (Thermo Fisher Scientific, A1217601) supplemented with primary hepatocyte thawing and plating supplements (Thermo Fisher Scientific, CM3000). Initial change of medium occurred 6 h after plating, with subsequent medium changes occurring every 24 h using William's E medium supplemented with primary hepatocyte maintenance supplements (Thermo Fisher Scientific, CM4000). For PhoenixBio primary human hepatocyte experiments, we obtained live human hepatocyte cultures (PXB-cells) from the provider in 96-well plates in DMEM without FBS. After arrival, hepatocytes were switched to hepatocyte growth medium (dHCGM, PhoenixBio) and maintained at 37 °C. Every 3–4 d, the culture medium was refreshed.

### Transfection

Cells were plated at 5,000–15,000 cells the day before transfection in a 96-well plate coated with poly-D-lysine (BD Biocoat). HEK293FT and HepG2 cells were transfected with Lipofectamine 2000 (Thermo Fisher Scientific) and GenJet HepG2 reagent (SignaGen, SL100489-HEPG2), respectively, according to manufacturer's specifications. For PASTE insertions, 100 ng of atgRNA guide-encoding plasmid, 250 ng of cargo plasmid, 50 ng of nicking guide-encoding plasmid and 375 ng of SpCas9-RT-P2A-Bxb1 complex-encoding plasmid were delivered to each well unless otherwise specified. For HITI insertions, 100 ng of guide-encoding plasmid, 250 ng of cargo plasmid and 75 ng of SpCas9 plasmid were delivered to cells. For HDR insertion of a large EGFP cargo, 100 ng of guide RNA, 200 ng of SpCas9 plasmid and 250 ng of insertion template plasmid were delivered to cells. Cells were replated 72 h later via limiting dilution to isolate clonal outgrowth in a 96-well plate for quantification of fluorescent colonies compared to PASTE. For HDR gene editing at the *EMX1* locus for non-dividing cell experiments, 300 ng of a single vector encoding the guide RNA, SpCas9 and HDR editing template was transfected, and cells were collected 72 h later for analysis by NGS. For PASTE experiments with hepatocytes, plasmids were transfected with standard Lipofectamine 3000 protocols with 400 ng of total plasmid. For Twin-PE knock-in experiments, transfection was performed as previously described[17] with Lipofectamine 2000 for *NOLC1* targeting or Lipofectamine 3000 for *ACTB* and *CCR5*.

### Plasmid electroporation

K562 and primary T cells were electroporated using a Lonza 4D-Nucleofector device (Lonza). The SF Cell Line 4D X Kit S (Lonza) was used for K562 cells, and the P3 Primary Cell 4D kit (Lonza) was used for the unstimulated primary T cells. Approximately $1.5 \times 10^6$ K562 cells were electroporated in a final volume of 20 µl in a 16-well nucleocuvette strip (Lonza). For the T cell experiments, $7.25 \times 10^6$ primary T cells were electroporated in a final volume of 100 µl in a cuvette.

For the single-vector and two-vector PASTE systems delivered to K562 cells, 900 ng of prime-Bxb1 complex-encoding plasmid or 800 ng of prime-encoding plasmid and 100 ng of Bxb1-encoding plasmid were electroporated, respectively. For both systems, 250 ng of cargo plasmid, 200 ng of atgRNA guide-encoding plasmid and 80 ng of RNA nicking guide-encoding plasmid were added.

For T cell electroporations, 990 ng of a guide vector expressing both the atgRNA and nicking guide, 875 ng of the EGFP-containing minicircle plasmid and 3,150 ng of the PASTE plasmid (SpCas9-RT-P2A-Bxb1) were electroporated.

Electroporations were performed according to the manufacturer's protocol, and after 72 h, the cells were collected for genomic DNA isolation and ddPCR quantification.

### Cloning of atgRNA efficiency screen library

atgRNA library members were computationally designed to cover corresponding ranges of PBS, RT and *attB* lengths. Each library member was also paired with a unique barcode and additional padding sequence after the poly(T) transcriptional terminator to maintain consistent oligonucleotide length. For each of the 12 spacer sequences in the library, corresponding library members were flanked by spacer-specific subpooling binding regions. The 10,580-member library was synthesized as a pool by Twist Biosciences and PCR amplified to generate 12 subpools. Each subpool was Golden Gate cloned into a corresponding backbone containing both the spacer sequence and a 200- to 300-nucleotide region for targeting. Each library was independently electroporated into Endura electrocompetent cells (Lucigen), plated on agar bioassay plates and collected the next day for protein purification.

### Pooled screening of atgRNA efficiency

The complete library was cotransfected with psPAX2 and pMD2.G with Lipofectamine 3000 (Thermo Fisher Scientific) to produce lentivirus for atgRNA library testing. Two days after transfection, supernatant containing virus was collected, filtered using 0.45-µm syringe filters and titered via spinfection, puromycin selection and Cell Titer Glow viability readout (Promega). After titer, the atgRNA viral library was used to infect 80 million HEK239FT cells at a multiplicity of infection of 0.3 to ensure single integration. After spinfection, cells were selected for 2 d with puromycin and allowed to expand and recover without drugs for an additional 2 d and were transfected with either PASTE constructs or Bxb1 integrase controls. Three days after transfection, cells were collected with the Quick gDNA midi kit (Zymo), and the corresponding library region was prepared for sequencing via PCR amplification. Prepared libraries were paired-end sequenced on an Illumina NextSeq 500.

### Computational analysis of the pooled atgRNA screen

Forward reads were trimmed to the corresponding barcode region to extract barcode sequences. Extracted barcodes were paired with corresponding targeted regions in the reverse read, which were trimmed to the region within 20 nucleotides of the putative *attB* region. To test for the presence or absence of editing, the region corresponding to the editing target was aligned to either the *attB* insertion outcome or the WT outcome, with reads aligning closer to the AttB insertion outcome being ranked as edited. Editing frequency was then taken as the ratio of edited to total reads for each barcode, with a psuedocount adjustment of 1.

### MLP modeling of atgRNA efficiency

Three different sequence-to-function models were considered for accurate prediction of atgRNA efficiency: simple linear/logistic regression, random forest classifier and MLP classifier. After the initial round of screening, we found that the MLP classifier performed the best and decided to move forward with a two-hidden-layer MLP model built in PyTorch. After initial optimization, the MLP classifier contained an input layer of 125 neurons, a first hidden layer of 512 neurons, a second hidden layer of 10 neurons and an output layer of 2 neurons. RELU was used as the activation function, and a dropout rate of 0.1 was applied for each layer. The output layer was transformed by a softmax function to predict probability for each class. To represent atgRNA as a vector, we considered simple one hot vector or *k*-mer breakdown of atgRNA. We varied the *k*-mer breakdown from 1 to 7 and found that breaking atgRNA into a short 3-nucleotide sequence (3-mer) was the most effective in training MLP. Padding was applied to atgRNA sequences that were shorter than 198 nucleotides, with '*N*' as the padding element to fulfill the input matrix to a uniform size. During the training of the MLP model, we varied the Adam optimizer's learning rate from 0.0001 to 0.01, batch size from 30 to 100 and epoch number from 10 to 100. We minimized the validation loss in a fivefold cross-validation algorithm with the cross entropy loss as the loss function and chose a learning rate of 0.001, epoch of 50 and batch size of 64 as the final training hyperparameters. ROC_AUC curve was performed using sklearn's roc_auc function. Codes to predict atgRNA efficiency and corresponding setup instructions are available at GitHub (https://github.com/abugoot-lab/atgRNA_rank).

### mRNA and synthetic guide electroporation

Before in vitro transcription, the DNA template was linearized by FastDigest MssI restriction enzyme (Thermo Fisher) and purified by QIAprep 2.0 Spin Miniprep columns (Qiagen). PASTE mRNA (SpCas9-RT-P2A-Bxb1) and *Bxb1* mRNA (NLS-Bxb1) were transcribed and poly(A) tailed using the HiScribe T7 ARCA mRNA kit (NEB, E2065S) with 50% supplement of 5-methyl-CTP and pseudo-UTP (Jena Biosciences), following the manufacturer's protocol. The mRNA was then purified using the MEGAclear transcription clean-up kit (Thermo Fisher, AM1908). For circularized *Bxb1* mRNA, in vitro transcription was conducted using the HiScribe T7 ARCA mRNA kit without modified nucleotides or the poly(A) tailing step. mRNA was subsequently circularized as previously reported[53] and cleaned again using the MEGAclear transcription clean-up kit. Additional full-length PASTE and *Bxb1* mRNAs were prepared by Trilink with CleanCap or'OMe Cleancap AG modifications and were fully substituted with $N^1$-methylpseudouridine. Chemically modified synthetic atgRNAs and nicking guides (Integrated DNA Technologies and Synthego) were provided by the corresponding parties. HEK293FT cells were electroporated using a Lonza 4D-Nucleofector device and the SF Cell Line 4D-Nucleofector X kit S (Lonza). For each sample, 4,000 ng of PASTE mRNA and 1,000 ng of

*Bxb1* mRNA were mixed with the designated amount of guide RNAs in a total volume of 15 µl of SF buffer solution. Cells ($2 \times 10^5$ per sample) were centrifuged at 100*g* for 10 min, resuspended in 5 µl of SF buffer solution and added to the 15-µl RNA solution. The 20-µl mixture was placed in one well of the cuvette strip and subjected to electroporation using the CM-130 program. Electroporated cells were resuspended in culture medium and incubated at 37 °C and 5% $CO_2$ for 72 h before analysis.

### Genomic DNA extraction and purification

DNA was collected from transfected cells by removal of medium, resuspension in 50 µl of QuickExtract (Lucigen) and incubation at 65 °C for 15 min, 68 °C for 15 min and 98 °C for 10 min. After thermocycling, lysates were purified via addition of 45 µl of AMPure magnetic beads (Beckman Coulter), mixing and two 75% ethanol wash steps. After purification, genomic DNA was eluted in 25 µl of water.

### Genome-editing quantification by ddPCR

To quantify PASTE and HITI editing efficiency by ddPCR, 24-µl solutions were prepared in a 96-well plate containing (1) 12 µl of 2× ddPCR Supermix for Probes (Bio-Rad), (2) primers for amplification of the integration junction at 250–900 nM, (3) FAM probe for detection of the integration junction amplicon at 250 nM, (4) 1.44 µl of RPP30 HEX reference mix (Bio-Rad), (5) 0.12 µl of FastDigest restriction enzyme for degradation of primer off-targets (Thermo Fisher) and (6) sample DNA at 1–10 ng $\mu l^{-1}$. All primers and probes used for ddPCR are listed in Supplementary Table 8. Twenty microliters of reaction mix was transferred to a Dg8 cartridge (Bio-Rad) and loaded into a QX2000 droplet generator (Bio-Rad). Forty-microliter droplets suspended in ddPCR droplet reader oil was transferred to a new 96-well plate and thermocycled according to manufacturer's specifications. The 96-well plate was then transferred to a QX200 droplet reader (Bio-Rad), and the generated data were analyzed using Quantasoft Analysis Pro to quantify DNA editing.

### Genome-editing quantification by targeted deep sequencing

To quantify integration of *attB/attP* pairs in the Bxb1 orthogonality assay and genome editing for prime editing and HDR integration at the *EMX1* locus, target regions were PCR amplified and analyzed by deep sequencing. Genomic DNA samples were isolated using 50 µl of QuickExtract (Lucigen) per well, and target regions were PCR amplified with NEBNext High-Fidelity 2× PCR master mix (NEB) based on the manufacturer's protocol. PCR amplicon primers are listed in Supplementary Table 9. Barcodes and adapters for Illumina sequencing were added in a subsequent PCR amplification. Amplicons were pooled and prepared for sequencing on a MiSeq (Illumina). Reads were demultiplexed and analyzed with appropriate pipelines. To analyze the Bxb1 orthogonality assay, *attP* barcodes were extracted and normalized to overall barcode counts, and WebLogos were generated with LogoMaker[63]. To analyze prime and HDR editing, amplicons were analyzed using custom scripts to analyze the relative number of reads with the inserted sequence. We also developed a three-primer NGS assay to quantify left junction integration using a common forward primer, a reverse primer to detect the unintegrated genomic locus and another reverse primer for detecting the insertion template. This assay was performed as above with each reverse primer at half concentration.

### RNA sequencing library preparation for analysis of transcriptome-wide off-targets

For analysis of Bxb1, prime and PASTE transcriptome effects, HEK293FT cells were transfected with corresponding vectors and collected after 3 d. Total RNA was purified using a Quick-RNA kit (Zymo), and mRNA was isolated from total RNA with a NEBNext poly(A) mRNA magnetic isolation module. Purified mRNA was prepared for NGS with a NEBNext ultra directional RNA library prep kit, and libraries were sequenced on an Illumina NextSeq instrument with a target of at least 5 million reads per library.

### RNA sequencing analysis pipeline for analysis of transcriptome-wide off-targets

RNA-sequencing samples were demultiplexed using custom scripts, checked for read quality with FastQC and aligned to the human GRCh38 genome index using the STAR[64] aligner package. Differential gene analysis was performed with edgeR, Limma and voom packages[65–67] to remove lowly expressed genes, normalize gene expression distributions and correct for non-heteroscedascity of count data when comparing between Bxb1, prime and PASTE effects on the transcriptome. Differentially expressed genes were considered significant if the absolute values of the differential gene expression was >0.585-fold and the *P* value was <0.05 after Benjamini–Hochberg correction. Volcano plots were generated to visualize the significance of differentially expressed transcripts in different libraries.

### Genome-wide off-target integration and integrase integration quantification by unique molecular identifier (UMI) Tn5 and NGS

To quantify the off-target integration of cargo payloads by PASTE and HITI throughout the human genome, single-cell clones were collected 3 d after transfection with QuickExtract (Lucigen) and purified using AMPure magnetic beads (Beckman Coulter) according to the manufacturer's protocol. Cellular genomic DNA was eluted in water and normalized to 6.25 ng $\mu l^{-1}$. A 2× Tn5 dialysis buffer was prepared with the following components according to Picelli et al. 2014[68]: 10 mM HEPES-KOH at pH 7.2, 0.2 M NaCl, 0.2 mM EDTA, 2 mM DTT, 0.2% Triton X-100 and 20% glycerol. Tn5 was assembled with equimolar preannealed mosaic-end double-stranded oligonucleotides by incubating the following components at room temperature for 1 h: 25 µl of 100 µM (final concentration) oligonucleotide mix in TE, 80 µl of 100% glycerol, 24 µl of 2× Tn5 dialysis buffer and 72 µl of Tn5 at an absorbance at 280 nm of 3.0. Normalized DNA (2.9 µl) was incubated with 0.1 µl of the Tn5 oligonucleotide mix and 0.75 µl of 5× Tris-HCl buffer (50 mM Tris-HCl, pH 8, and 25 mM $MgCl_2$) for 10 min at 55 °C. Then, 1.875 µl of the Tn5 transposition reaction was used as the template in a PCR reaction using SuperFi PCR master mix platinum (Thermo Fisher) according to the manufacturer's protocol. Next, 1 µl of the reaction was used as the template in an NGS reaction (see protocol below); UMI Tn5 primers for genome-wide off-target integration detection are listed in Supplementary Table 10. After NGS barcoding, all samples were diluted 1:1 and pooled; 20 µl of this pool was run on a 1% agarose gel, and a smear from 280 to 800 bp was extracted, purified and prepared for NGS on a MiSeq (Illumina).

To compare and quantify the integration efficiency of integrases, HEK293T cells were transfected with an atgRNA-expressing plasmid containing the *attB* site of the punitive integrase along with a minicircle and integrase-expressing plasmid; integration efficiency of the punitive integrase was measured as the integration of the minicircle into the atgRNA vector. To quantify this integration, the above UMI protocol was followed with different primer sets. The mosaic-end double-stranded oligonucleotides used in Tn5 preparation remained constant, as did the indexing reverse primer used in the SuperFi PCR mix and first-round NGS thermocycling steps. The forward primers for these thermocycling steps were changed for primers with homology for the atgRNA acceptor plasmid. Integrase reporter primers can be found in Supplementary Table 10.

### Quantification of in vivo editing efficiency by three-primer NGS

To quantify in vivo PASTE editing efficiency by three-primer NGS, a 5-µl reaction was prepared containing 2.5 µl of NebNext PCR mix (New England Biolabs), 2 µl of water/primer mix and 0.5 µl of mouse liver DNA normalized to 40 ng $\mu l^{-1}$ (purified as described above). For left (*attL*) junction analysis, a pool of forward primers binding upstream of the endogenous edited site was paired with two reverse primers, one binding downstream of the endogenous edited site and one binding in the PASTE-integrated minicircle. For right (*attR*) junction analysis, a single reverse primer binding downstream of the endogenous edited site was paired with two forward primer pools, one binding upstream of the endogenous edited site and one binding in the PASTE-integrated minicircle (Supplementary Table 9). To avoid PCR bias, primers were positioned to ensure both amplicons generated in the subsequent PCR reaction were of equivalent length (±5 bp). The final concentrations of reverse and forward primers in solution were equivalent (1 μM total). The first-round PCR reaction was run for 18 cycles and barcoded for an additional 20 cycles according to the manufacturer's protocol. Samples were then prepared for NGS on a MiSeq (Illumina) as described above.

### Validation of the three-primer NGS method with PCR standards

To validate the accuracy of our three-primer NGS method for quantifying PASTE integration efficiency, a series of PCR standards was prepared. An amplicon with the sequence of the unedited endogenous site and an amplicon with the sequence of the PASTE-edited endogenous site were mixed in a dilution series of 100:0, 25:75, 50:50, 75:25 and 0:100. Three-primer NGS was performed on these PCR standards as described above, and the measured 'editing efficiency' of the standards was compared next to the standard's amplicon composition (Extended Data Fig. 10a). The measured editing efficiency of the standards had strong concordance with the amplicon composition of the input standards, validating the accuracy of the three-primer NGS method for quantifying PASTE integration efficiency.

### Computational identification of Bxb1 and Cas9 off-targets

To identify potential off-target sites for Bxb1 integration or Cas9 cleavage, similar sequences were identified in the human genome using either BLAST[69] for similar *attP* sequences or Cas9 off-target prediction algorithms[70]. To validate successful amplicon generation by primer sets, positive-control off-target amplicons were ordered as oligonucleotides, annealed and tested by ddPCR. Off-target sites are listed in Supplementary Table 11.

For genome-wide characterization of off-targets, reads were filtered for reads carrying the cargo sequence and trimmed to the genomic junction. Reads were then BWA aligned to the human genome (GRCh38/hg38) and filtered for alignment lengths of at least 30 bp. Filtered reads were then analyzed for coverage and plotted.

### Imaging

For sample preparation for imaging, coverslips were placed at the bottom of a 24-well plate before plating HEK293FT cells. After transfection at ~70% confluency and an incubation period of 3 d, the medium was removed, and the wells were washed with 1 ml of PBS (pH 7.4; Thermo Fisher Scientific). Cells were fixed with 500 μl of 4% Pierce formaldehyde (Thermo Fisher Scientific) for 30 min. Additional washing with 1 ml of PBS (pH 7.4) was performed three times. If no immunostaining was to be performed, slides were processed to be mounted.

If immunostaining was to be performed, the cells were blocked in 1 ml of 2.5% goat serum (Cell Signaling Technology) and 0.1% Triton X-100 (Sigma-Aldrich) for 1 h at room temperature. For the primary stain, the primary antibodies were mixed with 1.25% goat serum, and 300 μl was added per well according to the following dilutions: 1:1,500 for anti-ACTB (NB600-501SS, NovusBio), 1:200 for anti-SRRM2 (NBP2-55697, NovusBio), 1:200 for anti-NOLC1 (11815-1-AP, ProteinTech) and 1:200 for anti-Lamin B1 (12987-1-AP, ProteinTech). After shaking overnight at 4 °C, the wells were washed three times with 1 ml of PBS (pH 7.4). For the secondary staining, a 1:1,000 dilution of secondary antibody, either goat anti-mouse IgG Alexa Fluor 568 (Thermo Fisher Scientific, A-11004) or goat anti-rabbit IgG Alexa Fluor 647 (Thermo Fisher Scientific, A21244), was mixed with 1.25% goat serum. After 1 h at room temperature, cells were washed with PBS three times, and slides were mounted.

To mount the slides, a drop of ProLong Gold Antifade Mountant with DAPI (Thermo Fisher Scientific) was placed on the top of the slide, and the coverslips were removed from the 24-well plate and inverted onto the drop. The coverslips were left to dry for 24 h protected from light at room temperature and sealed with nail polish. For acquisition of images, a laser scanning confocal microscope (Zeiss, LSM900) was used with a ×40 oil objective and three different filter sets for visualizing EGFP, DAPI and the immunofluorescence stain (either 568 nm or 647 nm).

### AAV production and transduction

To produce AAV vectors for delivery of PASTE cargo, HEK293FT cells were transfected in T25 flasks using Lipofectamine 3000 (Thermo Fisher Scientific) with 1.6 μg of GFP AAV cargo plasmid, 1.96 μg of AAV8 capsid vector and 4.13 μg of AAV helper pAdDeltaF6 plasmid (Addgene, 112867) per T25 flask according to manufacturer's protocol. Two days after transfection, the medium containing the loaded viral vector was filtered using a 0.45-μm filter (Sigma-Aldrich), and the final product was stored at −80 °C. One day after the transfection of PASTE components (PE2, Bxb1, nicking guide and atgRNA) into HEK293FT cells, AAVs containing the GFP cargo template were introduced directly into the cells according to the indicated volumes. Three days after the transduction, the cells were collected, and ddPCR readout was performed.

### AdV production and transduction

AdV vectors were cloned using the AdEasy-1 system obtained from Addgene. Briefly, SpCas9-RT-P2A-Blast, Bxb1 and guide RNAs and an EGFP cargo gene were cloned into separate AdV template backbones and recombined to add the full AdV genome with the AdEasy-1 plasmid in BJ5183 *E. coli* cells. These recombined plasmids were sent to Vector BioLabs for commercial production. Additional AdV vectors were produced for in vivo experiments by the University of Massachusetts Medical School Viral Vector Core, as previously described[71–73].

For the EGFP cargo vector, EGFP cargo Adv vector was added at $6.7 \times 10^6$ plaque-forming units (p.f.u.) per well of a 96-well plate of HEK293FT cells and $1.3 \times 10^6$ p.f.u. per well of a 96-well plate of HepG2 cells. For experiments with the three-vector AdV delivery of all PASTE components, we used $8 \times 10^5$ p.f.u. of each viral vector per well of a 96-well plate of HEK293FT cells. For three-vector AdV delivery on HepG2 cells, we used $40 \times 10^6$ p.f.u. of the EGFP cargo vector, $10 \times 10^6$ p.f.u. of the SpCas9-RT-P2A-Blast vector and $20 \times 10^6$ p.f.u. of the Bxb1 and guides vector per well of a 96-well plate. For three-vector AdV delivery to PhoenixBio primary human hepatocytes (PXB-cells), we used the vector amounts listed in the figure and transduced for 3 d before collection for ddPCR analysis.

### Quantification of protein expression

Three days after the transfection of HepG2 cells, the Nano-Glo HiBiT lytic detection system (Promega) was used for the quantification of the HiBiT-tagged proteins SERPINA1 and CPS1 in cell lysates or medium. For the preparation of the Nano-Glo HiBiT lytic reagent, the Nano-Glo HiBit lytic buffer (Promega) was mixed with Nano-Glo HiBiT lytic substrate (Promega) and the LgBiT protein (Promega) according to manufacturer's protocol. The volume of Nano-Glo HiBiT lytic reagent added was equal to the culture medium present in each well, and the samples were placed on an orbital shaker at 600 r.p.m. for 3 min. After incubation for 10 min at room temperature, the readout was performed with 125 gain and 2 s integration time using a plate reader (Biotek Synergy Neo 2). The control background was subtracted from the final measurements.

### Computational discovery of serine integrases

Prokaryotic genomic and metagenomic sequences were retrieved from various public databases and datasets, including NCBI, European Nucleotide Archive (ENA), Ensembl, MetaSUB, MGnify and JGI. Protein-coding genes were predicted with Prodigal v2.6.3 (ref. [74]).

Protein sequences were scanned for large serine integrase domains with hmmsearch (HMMER v3.3.2)[75] using Pfam models PF00239, PF07508 and PF13408 with model-specific gathering cutoffs. Protein sequences not containing at least a resolvase and recombinase domain were discarded, and the remaining sequences were marked as putative large serine integrases. The source contigs of these putative integrases were passed to VirSorter v1.0.6 (ref. [76]) for prophage prediction. Contigs that were predicted to have a putative integrase completely embedded in a prophage region were kept, and the 1,000 bp around the predicted prophage boundaries were searched for *k*-mer matches of 2–18 bp with the 100 bp around the predicted integrase. Matching *k*-mer sites were then expanded to 50 bp and scanned for inverted repeats. Sites with a high number of dinucleotide inverted repeats (based on an experimentally derived cutoff) were nominated as putative attachment sites. To expand the set of integrases and improve attachment site prediction accuracy, another mining approach was applied to all sequences with a species-level taxonomic annotation. All of the assemblies with a predicted integrase were paired with an assembly from the same species without a predicted integrase. MGEfinder v1.0.6 (ref. [77]) was used on each pair to predict mobile genetic elements. Predicted integrases that were completely embedded in a predicted MGE region were kept, and the same attachment site prediction algorithm was applied to their contigs with a reduced search of 30 bp. The two sets of integrases were pooled, and the attachment sites predicted using the MGEfinder method took precedence in the case of multiple predictions. The pooled set of integrases was then used to search the NCBI CDD using RPS-BLAST with a E value threshold of 0.001, and integrases without a hit to a serine recombinase resolvase domain were discarded. The remaining sequences were clustered with CD-HIT v4.8.1 using the −c 0.7 -s 0.8 options (70% sequence identity and 80% coverage of the shorter sequence). The sequences were aligned with MUSCLE v5.0.1278, and sequences not aligning to the integrase catalytic residues were discarded. The remaining sequences were used to generate a phylogenetic tree with FastTree v2.1.11 using the LG + CAT substitution model. Clades were chosen with manual inspection, and domain architectures were visualized with HHpred.

### Western blotting analysis of protein levels

Cells were lysed in cell lysis buffer (Invitrogen). Equal volumes of cell lysate were run on Mini-PROTEAN TGX stain-free precast gels (Bio-Rad) and transferred to nitrocellulose membranes using the iBlot 2 dry blotting system (Thermo Scientific). Nonspecific antigen binding was blocked with LICOR Intercept (PBS) blocking buffer for 1 h at room temperature. Membranes were then incubated with primary antibodies (β-actin antibody (4967S) with GAPDH antibody (97166S) or lamin B1 antibody (12586S) with GAPDH antibody (97166S)) from Cell Signaling Technology. Antibodies were diluted at 1:1,000 in Intercept (PBS) blocking buffer with 0.2% Tween-20, and the membranes were incubated for 16 h at 4 °C. Membranes were washed four times for 5 min each with PBS + 0.2% Tween-20 and further incubated with LICOR donkey anti-rabbit IgG polyclonal antibody (IRDye 800CW) and goat anti-mouse IgG polyclonal antibody (IRDye 680RD) diluted 1:15,000 in Intercept (PBS) blocking buffer with 0.2% Tween-20. The membranes were incubated for 1 h at room temperature followed by four 5-min washes in PBS + 0.2% Tween-20. The membranes were imaged using an Odyssey scanner (LICOR Biosciences) and analyzed by band densitometry using ImageJ.

### In vivo injections of PASTE AdV

PASTE AdV preparations were prepared for the two in vivo conditions: (1) $8 \times 10^{10}$ p.f.u. of the EGFP cargo vector (University of Massachusetts Medical), $2 \times 10^{8}$ p.f.u. of the SpCas9-RT-P2A-Blast vector (Vector Biolabs) and $2 \times 10^{10}$ p.f.u. of the Bxb1 and guides vector (University of Massachusetts Medical) and (2) $8 \times 10^{10}$ p.f.u. of the EGFP cargo vector (University of Massachusetts Medical) and $2 \times 10^{10}$ p.f.u. of the Bxb1 and guides vector (University of Massachusetts Medical). These preparations were constituted in 150 μl of PBS and shipped to Yecuris for injections in the liver-humanized FRG knockout mouse model. Mice were injected at ~5.5 months of age via retroorbital injection. All mice enrolled in the study were removed from 2-(2-nitro-4-trifluromethylbenzoyl)-1,3-cyclohexanedione (nitisinone; NTBC) for ≥20 d and Trimethoprim-sulfamethoxazole (TMP-SMX) for ≥3 d before dosing to reduce the contribution of mouse hepatocytes. Liver humanization was evaluated ≤7 d before the start of the study by human serum albumin enzyme-linked immunosorbent assay (Bethyl Laboratory, E90-134). All mice were weighed before the start of the study and evenly grouped based on their human serum albumin concentration and body weights. One day after dosing, the mice were treated with NTBC for 3 d and then continued the standard water formulation containing dextrose and vitamin C for the duration of the study. Mice were maintained at the Yecuris Corporation-affiliated Institutional Animal Care and Use Committee (IACUC)-accredited facility. General procedures for animal care and housing were as described in the Guide for the Care and Use of Laboratory Animals, National Research Council, Yecuris IACUC Policy and Yecuris General Mouse Handling Care and Euthanasia. Cages were changed every 2 weeks, and the testing facility was sanitized weekly. Animal studies were performed in accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health (NIH). The protocols were approved by the IACUC at the Massachusetts Institute of Technology (protocol number 0919-065-22) and Yecuris Corporation IACUC Policy.

At ~3 weeks after injection, the chimeric livers were collected and snap frozen. Liver pieces were sectioned from different regions, and genomic DNA was purified using the Qiagen DNeasy Blood and Tissue kit. Genomic DNA was then subjected to a three-primer sequencing assay for quantifying the left junction integration of the AdV template at the human-specific *ACTB* locus.

### Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Raw reads for RNA sequencing and the atgRNA efficiency screen are available at Sequence Read Archive under BioProject accession number PRJNA700575 (ref. [78]). Expression plasmids are available from Addgene at https://www.addgene.org/browse/article/28223250/ under UBMTA. The human genome GRCh38 can be accessed at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/. Source data are provided with this paper.

## Code availability

Code to predict atgRNA efficiency and support information is available at https://github.com/abugoot-lab/atgRNA_rank[79].

## References

63. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).
64. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
65. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
66. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
67. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

68. Picelli, S., Åsa K., Björklund, A.K., Reinius, B., Sagasser, S. Winberg, G. and Sandberg, R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. 24, 2033–2040 (2014).
69. Johnson, M. et al. NCBI BLAST: a better web interface. *Nucleic Acids Res*. **36**, W5–W9 (2008).
70. Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol*. **31**, 827–832 (2013).
71. Sena-Esteves, M. & Gao, G. Introducing genes into mammalian cells: viral vectors. *Cold Spring Harb. Protoc*. **2020**, 095513 (2020).
72. Su, Q., Sena-Esteves, M. & Gao, G. Release of the cloned recombinant adenovirus genome for rescue and expansion. *Cold Spring Harb. Protoc*. **2019**, https://doi.org/10.1101/pdb.prot095539 (2019).
73. Su, Q., Sena-Esteves, M. & Gao, G. Purification of the recombinant adenovirus by cesium chloride gradient centrifugation. *Cold Spring Harb. Protoc*. **2019**, https://doi.org/10.1101/pdb.prot095547 (2019).
74. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
75. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol*. **7**, e1002195 (2011).
76. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).
77. Durrant, M. G., Li, M. M., Siranosian, B. A., Montgomery, S. B. & Bhatt, A. S. A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. *Cell Host Microbe* **28**, 140–153 (2020).
78. Yarnall, M. T. N. et al. Genome insertion of large sequences without double-strand DNA cleavage using CRISPR-directed integrases. *SRA* https://www.ncbi.nlm.nih.gov/bioproject/PRJNA700575/ (2022).
79. Yarnall, M. T. N. et al. Genome insertion of large sequences without double-strand DNA cleavage using CRISPR-directed integrases. *GitHub* https://github.com/abugoot-lab/atgRNA_rank (2022).
80. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol*. **3**, Article3 (2004).
81. McCarthy, D. J. & Smyth, G. K. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics* **25**, 765–771 (2009).

## Acknowledgements

We would like to thank B. Desimone, F. Chen, J. Joung, A. Serj-Hansen, G. Feng, J. Wilde, M. Calos, T. Aida, Y. Cha and M. Mittens for helpful discussions, E.V. Koonin and K. Makarova for helpful discussions with integrase discovery and annotation, P. Reginato, D. Weston and E. Boyden for MiSeq instrumentation, S. Jacobs and A. Ainbinder for ddPCR instrumentations, S. Bhatia and S. March Riera for hepatocyte assistance, G. Paradis and M. Griffin for flow cytometry assistance and J. Crittenden for editing the manuscript. L.V. is supported by a Swiss National Science Foundation Postdoc Mobility Fellowship. O.O.A. and J.S.G. are supported by NIH grants 1R21-AI149694, R01-EB031957 and R56-HG011857, The McGovern Institute Neurotechnology program, the K. Lisa Yang and Hock E. Tan Center for Molecular Therapeutics in Neuroscience, G. Harold & Leila Y. Mathers Charitable Foundation, MIT John W. Jarve (1978) Seed Fund for Science Innovation, Impetus Grants, Cystic Fibrosis Foundation Pioneer Grant, Google Ventures, FastGrants, the Harvey Family Foundation and the McGovern Institute. S.K.G. was supported by the Intramural Research Program of the National Library of Medicine, NIH.

## Author contributions

O.O.A. and J.S.G. conceived the study. O.O.A. and J.S.G. designed and participated in all experiments. M.T.N.Y., E.I.I. and C.S.-U. led many of the experiments and assay readouts. R.N.K. helped with cell culture, cloning, plasmid sequencing, NGS and in vivo experiments. M.T.N.Y. and C.S.-U. helped with ddPCR, sequencing experiments and cloning. L.V. helped with various PASTE editing experiments and characterization of integrases. W.Z. synthesized mRNA and performed the electroporation experiments. J.L. and S.K.G. performed the computational mining to uncover integrases and annotated these new systems. K.J. performed the ML modeling of the pooled atgRNA screening and developed a guide design software package. N.R., L.Z., K.H., J.A.W, A.P.K., A.E.Z. and C.A.V. synthesized synthetic guides and advised on synthetic RNA experiments. J.M.H. and A.U. provided select mRNA constructs and advised on mRNA experiments. H.M., J.X. and G.G. produced AAV and AdV. S.K.D., Y.M. and D.R.R. provided primary human hepatocytes and advice for in vivo experiments with humanized mouse models. L.F. and G.B. provided humanized liver mice, managed in vivo injections and collections and advised on the in vivo aspects of the project. O.O.A. and J.S.G. wrote the manuscript with help from all authors.

## Competing interests

O.O.A., J.S.G., J.L., L.V. and K.J. are co-inventors on patent applications filed by Massachusetts Institute of Technology relating to work in this manuscript. O.O.A. and J.S.G. are cofounders of Sherlock Biosciences, Proof Diagnostics, Moment Biosciences and Tome Biosciences. O.O.A. and J.S.G. were advisors for Beam Therapeutics during the course of this project. K.H., J.A.W., A.P.K. and A.E.Z. are employees and shareholders of Synthego. S.K.D., Y.M. and D.R.R. are employees of PhoenixBio. L.F. and G.B. are employees of Yecuris Corporation. N.R., L.Z. and C.A.V. are employees of Integrated DNA Technologies. The remaining authors declare no competing interest.

## Additional information

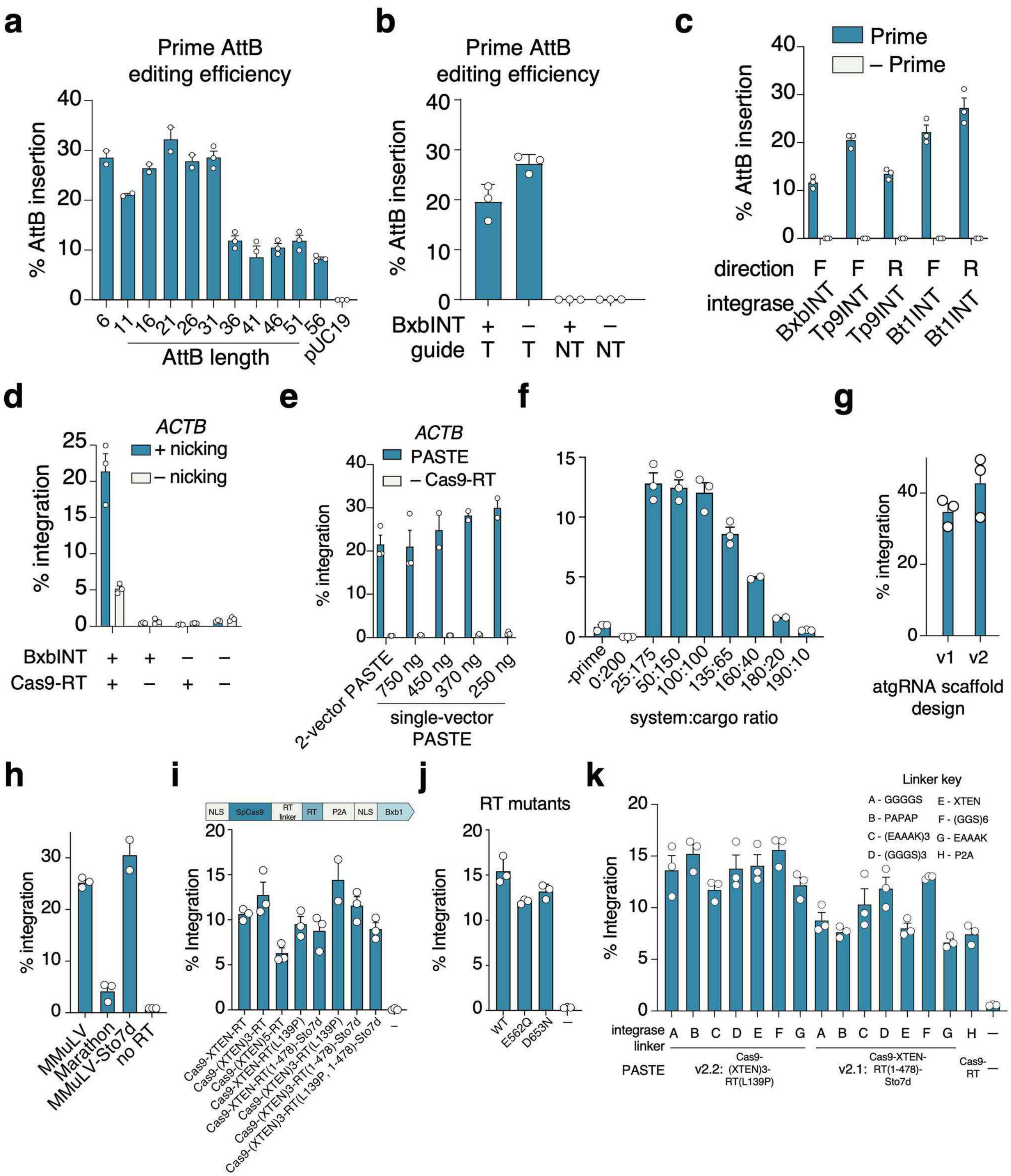
**Extended data** is available for this paper at https://doi.org/10.1038/s41587-022-01527-4.

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01527-4.

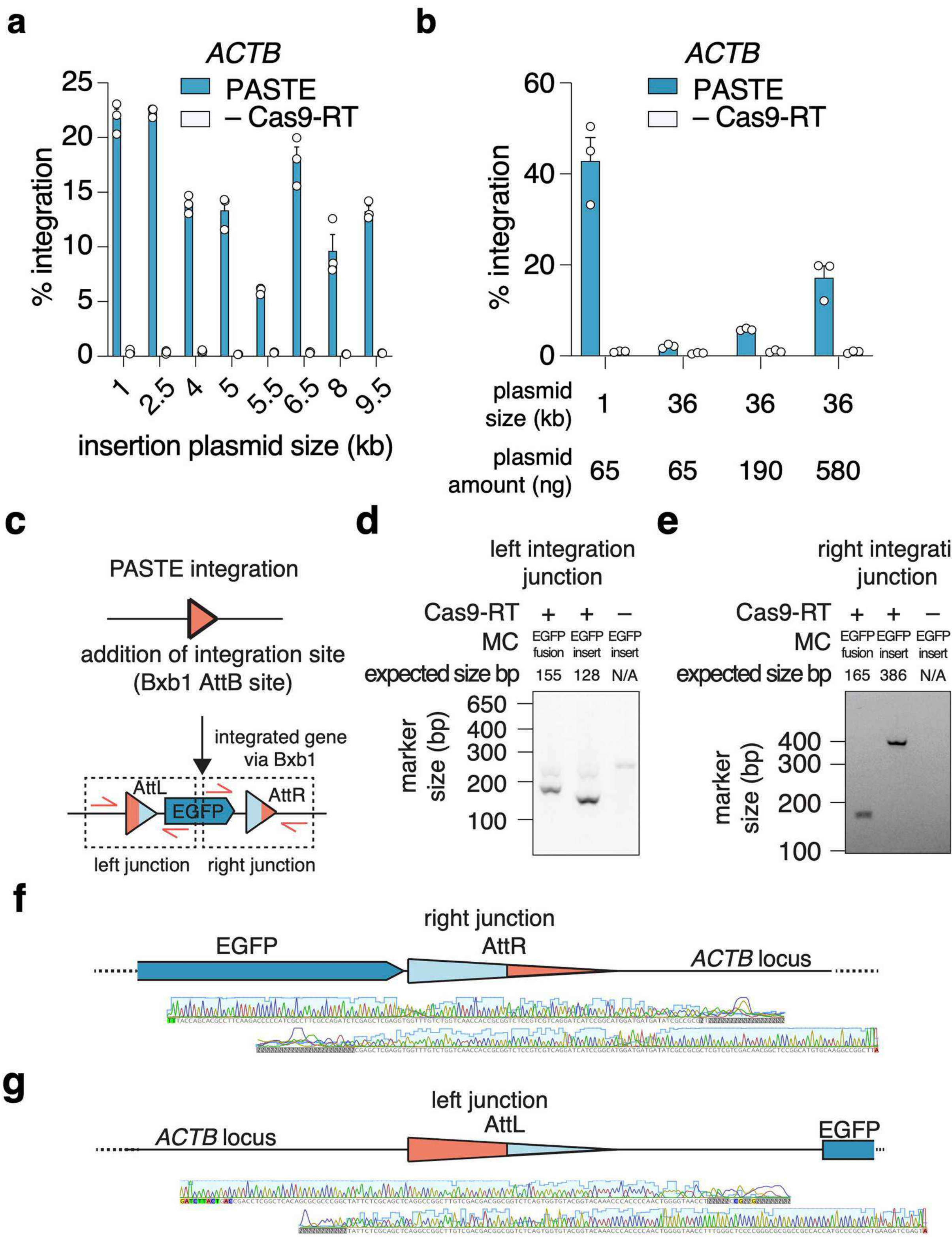
**Correspondence and requests for materials** should be addressed to Omar O. Abudayyeh or Jonathan S. Gootenberg.

**Peer review information** *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

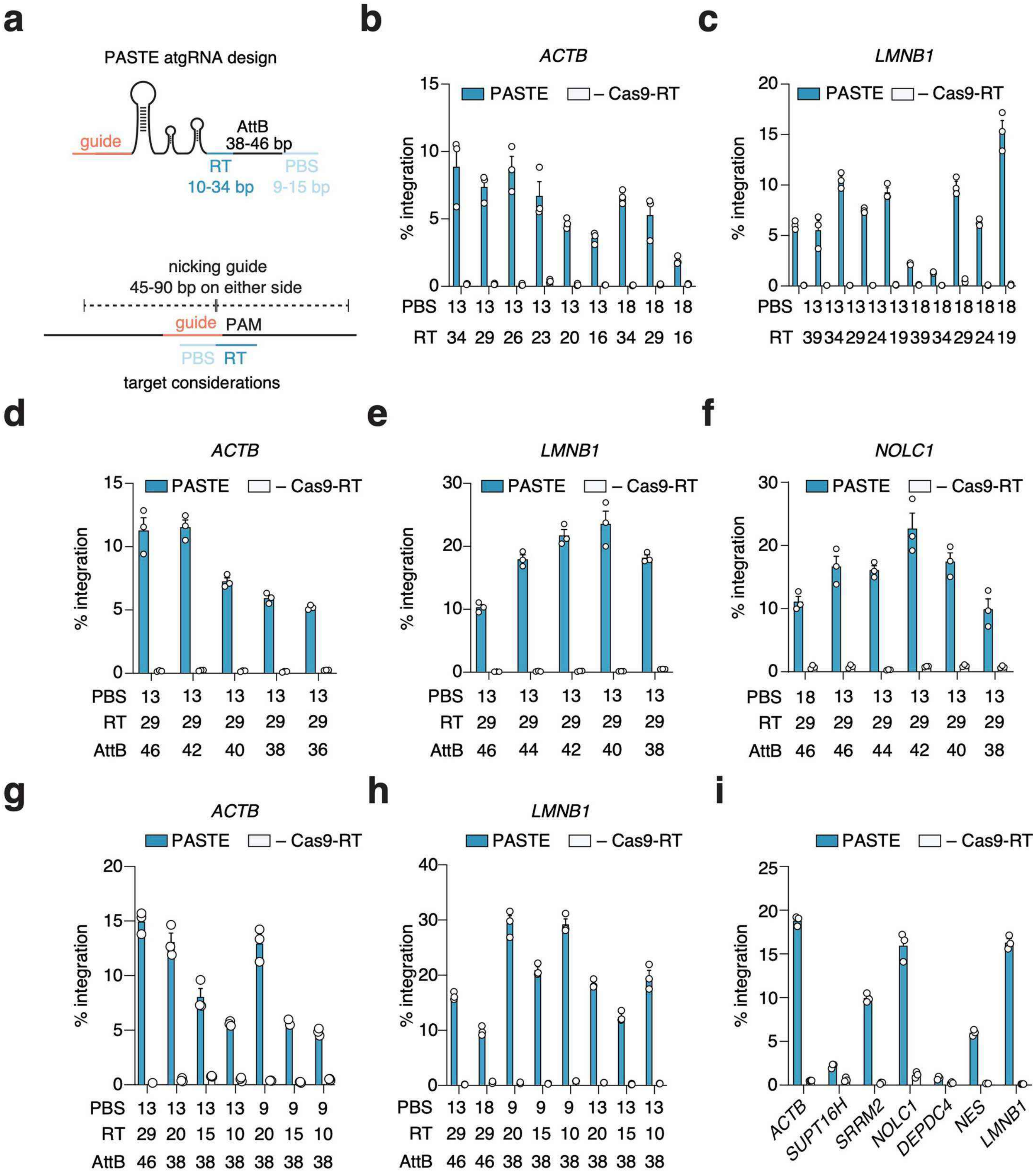
**Reprints and permissions information** is available at www.nature.com/reprints.



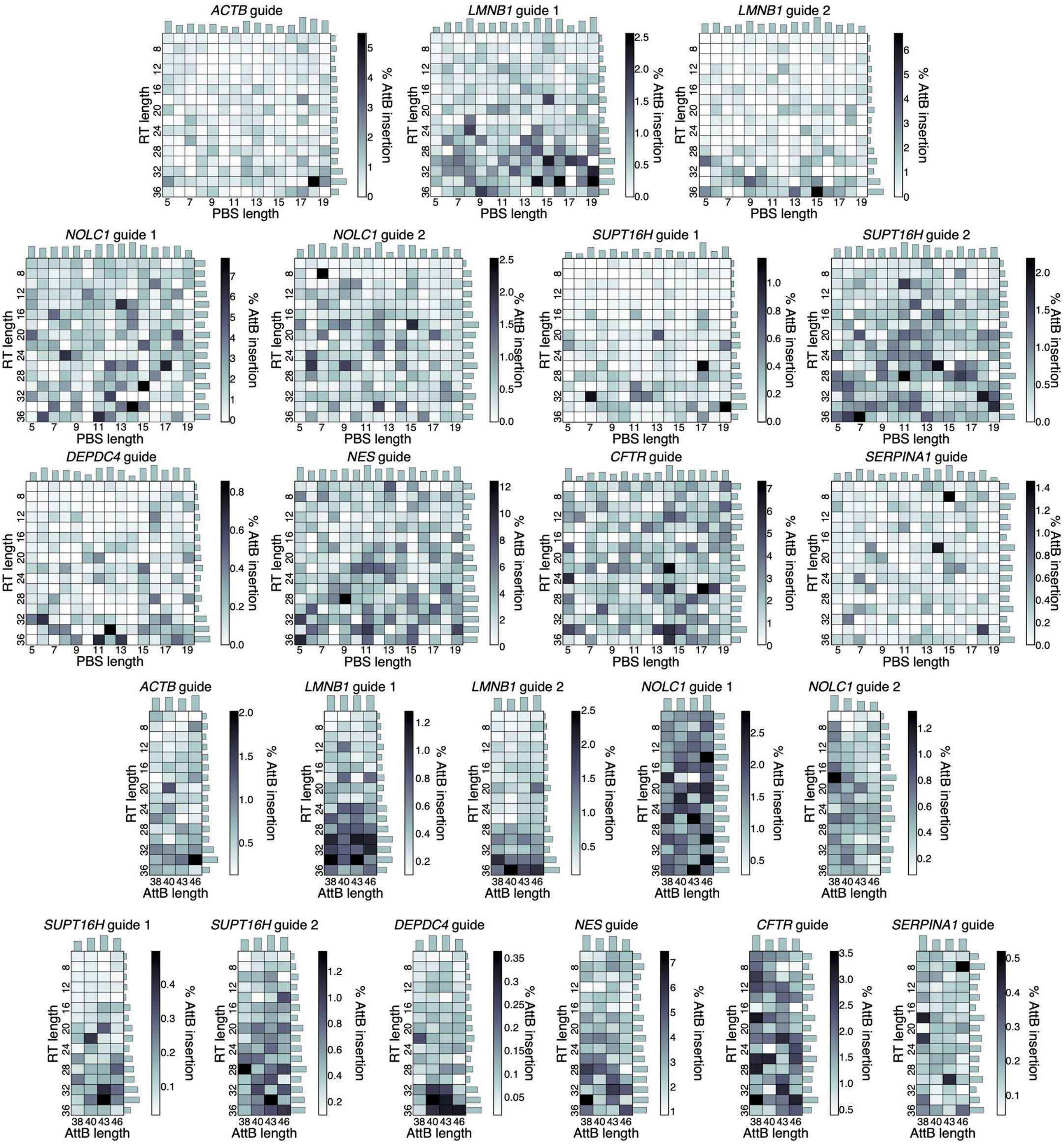
**Extended Data Fig. 1 | Evaluation of prime integration activity for diverse *attB* sequences and optimization of PASTE editing through dosage and mutagenesis. a)** Prime editing efficiency for the insertion of different length BxbINT *attB* sites at *ACTB*. Data are mean (n = 2 or 3) ± s.e.m. **b)** Prime editing efficiency for this insertion of a BxbINT *attB* site at *ACTB* with targeting and non-targeting guides. Data are mean (n = 3) ± s.e.m. **c)** Prime editing efficiency for the insertion of different integrases' *attB* sites at *ACTB*. Both orientations of landing sites are profiled (F, forward; R, reverse). Data are mean (n = 3) ± s.e.m. **d)** PASTE editing efficiency for the insertion of EGFP at *ACTB* with and without a nicking guide. Data are mean (n = 3) ± s.e.m. **e)** PASTE integration efficiency of EGFP at *ACTB* measured with different doses of a single-vector delivery of components. Data are mean (n = 2 or 3) ± s.e.m. **f)** PASTE integration efficiency of EGFP at *ACTB* measured with different ratios of a single-vector delivery of components to the EGFP template vector. Data are mean (n = 3) ± s.e.m. **g)** PASTE efficiency at the *ACTB* target compared between atgRNAs containing either the v1 or v2 scaffold designs. Data are mean (n = 3) ± s.e.m. **h)** PASTE integration efficiency of EGFP at *ACTB* with different RT domain fusions. Data are mean (n = 2 or 3) ± s.e.m. **i)** PASTE integration efficiency of EGFP at *ACTB* with different RT domain fusions and linkers. Data are mean (n = 2 or 3) ± s.e.m. **j)** PASTE integration efficiency of EGFP at *ACTB* with mutant RT domains. Data are mean (n = 3) ± s.e.m. **k)** Optimization of PASTE constructs with a panel of linkers and RT modifications for Gluc integration at the *ACTB* locus using atgRNAs with the v2 scaffold. Data are mean (n = 3) ± s.e.m.



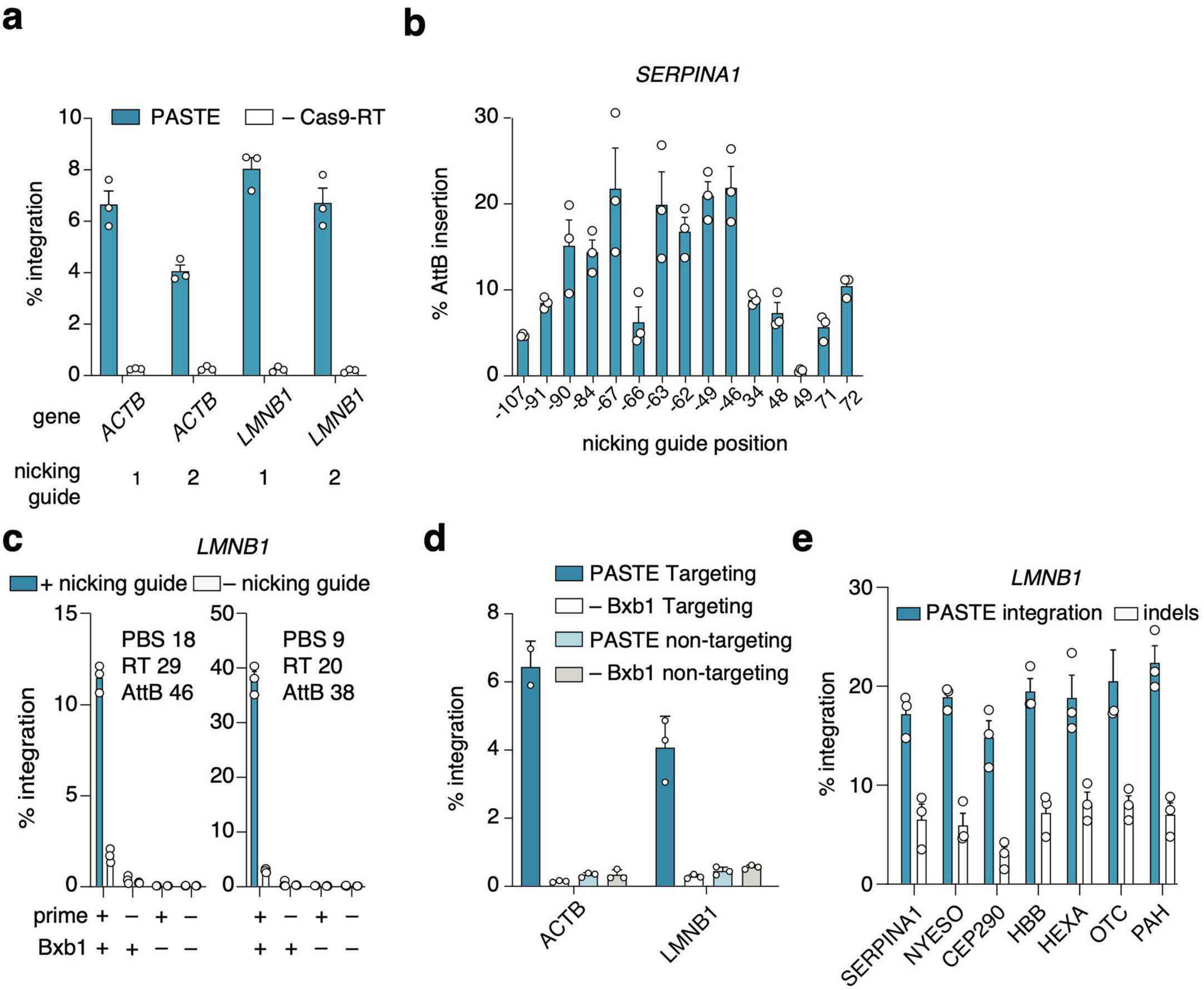
**Extended Data Fig. 2 | Characterization of PASTE payload sizes and integration junctions. a**) PASTE integration efficiency at the *ACTB* locus of varying sized cargos transfected at a fixed DNA amount and variable molar ratio. **b**) PASTE integration efficiency at the *ACTB* locus of varying sized cargos transfected at a variable DNA amounts. **c**) Schematic of PASTE integration, including resulting *attB* and *attL* sites that are generated and PCR primers for assaying the integration junctions. **d**) PCR and gel electrophoresis readout of left integration junction from PASTE insertion of GFP at the *ACTB* locus. Insertion is analyzed for in-frame and out-of-frame GFP integration experiments as well as for a no prime control. Expected sizes of the PCR fragments are shown using the primers shown in the schematic in subpanel A. e) PCR and gel electrophoresis readout of right integration junction from PASTE insertion of GFP at the *ACTB* locus. Insertion is analyzed for in-frame and out-of-frame GFP integration experiments as well as for a no prime control. Expected sizes of the PCR fragments are shown using the primers shown in the schematic in subpanel A. f) Sanger sequencing shown for the right integration junction for an in-frame fusion of GFP via PASTE to the N-terminus of *ACTB*. g) Sanger sequencing shown for the left integration junction for an in-frame fusion of GFP via PASTE to the N-terminus of β-actin. Data are mean (n = 3) ± s.e.m.



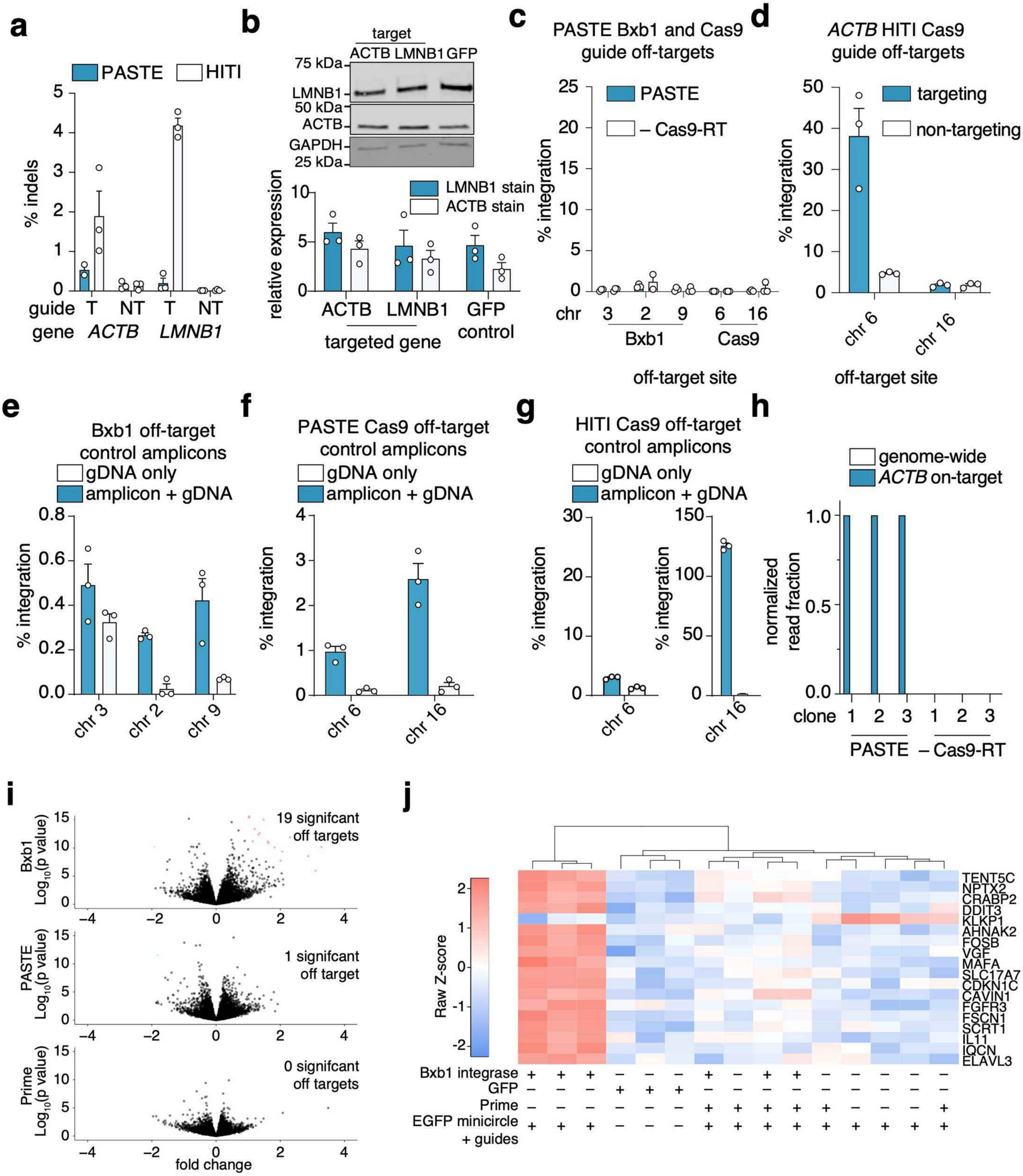
**Extended Data Fig. 3 | Validation of design rules for efficient PASTE insertion at endogenous genomic loci. a**) Schematic of various parameters that affect PASTE integration of ~1 kb GFP insert. On the atgRNA, the PBS, RT, and *attB* lengths can alter the efficiency of AttB insertion. Nicking guide selection also affects overall gene integration efficiency. **b**) The impact of PBS and RT length on PASTE integration of *GFP* at the *ACTB* locus. **c**) The impact of PBS and RT length on PASTE integration of GFP at the *LMNB1* locus. **d**) The impact of *attB* length on PASTE integration of *GFP* at the *ACTB* locus. **e**) The impact of *attB* length on PASTE integration of GFP at the *LMNB1* locus. **f**) The impact of *attB* length on PASTE integration of *GFP* at the *NOLC1* locus. **g**) The impact of minimal PBS, RT, and *attB* lengths on PASTE integration efficiency of *GFP* at the *ACTB* locus. h) The impact of minimal PBS, RT, and *attB* lengths on PASTE integration efficiency of *GFP* at the *LMNB1* locus. i) PASTE integration efficiency of EGFP at varying endogenous loci. Data are mean (n = 3) ± s.e.m.



**Extended Data Fig. 4 | Heatmaps depicting the effect of PBS, RT, and *attB* lengths on atgRNA efficiency of attachment site insertion from high-throughput pooled screening of 10,580 guides targeting a variety of loci.** Bar charts indicating normalized summation across relevant PBS, RT, or *attB* parameter axes are shown on heatmap sides.

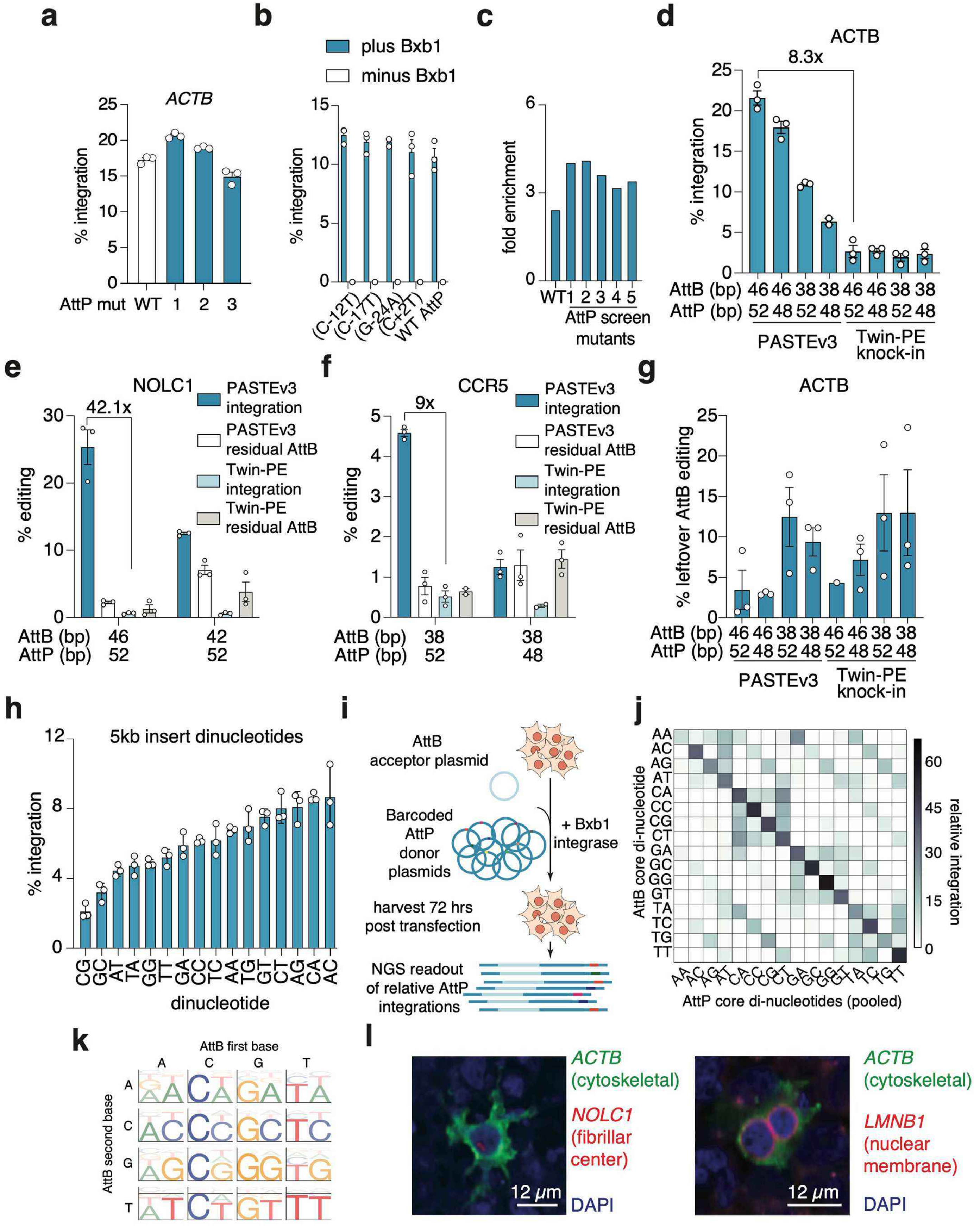


**Extended Data Fig. 5 | Effect of nicking guides on insertion of diverse cargos.** **a**) PASTE insertion efficiency at *ACTB* and *LMNB1* loci with two different nicking guide designs. **b**) Attachment site insertion at the *SERPINA1* locus with a panel of different nicking guides at varying distances. **c**) Effect of nicking guides on PASTE integration efficiency at the *LMNB1* locus with two different atgRNA designs. **d**) PASTE integration efficiency at *ACTB* and *LMNB1* with target and non-targeting spacers and matched atgRNAs with and without BxbINT expression. e) Integration of a panel of different gene cargo at *LMNB1* locus via PASTE. Data are mean (n = 3) ± s.e.m.



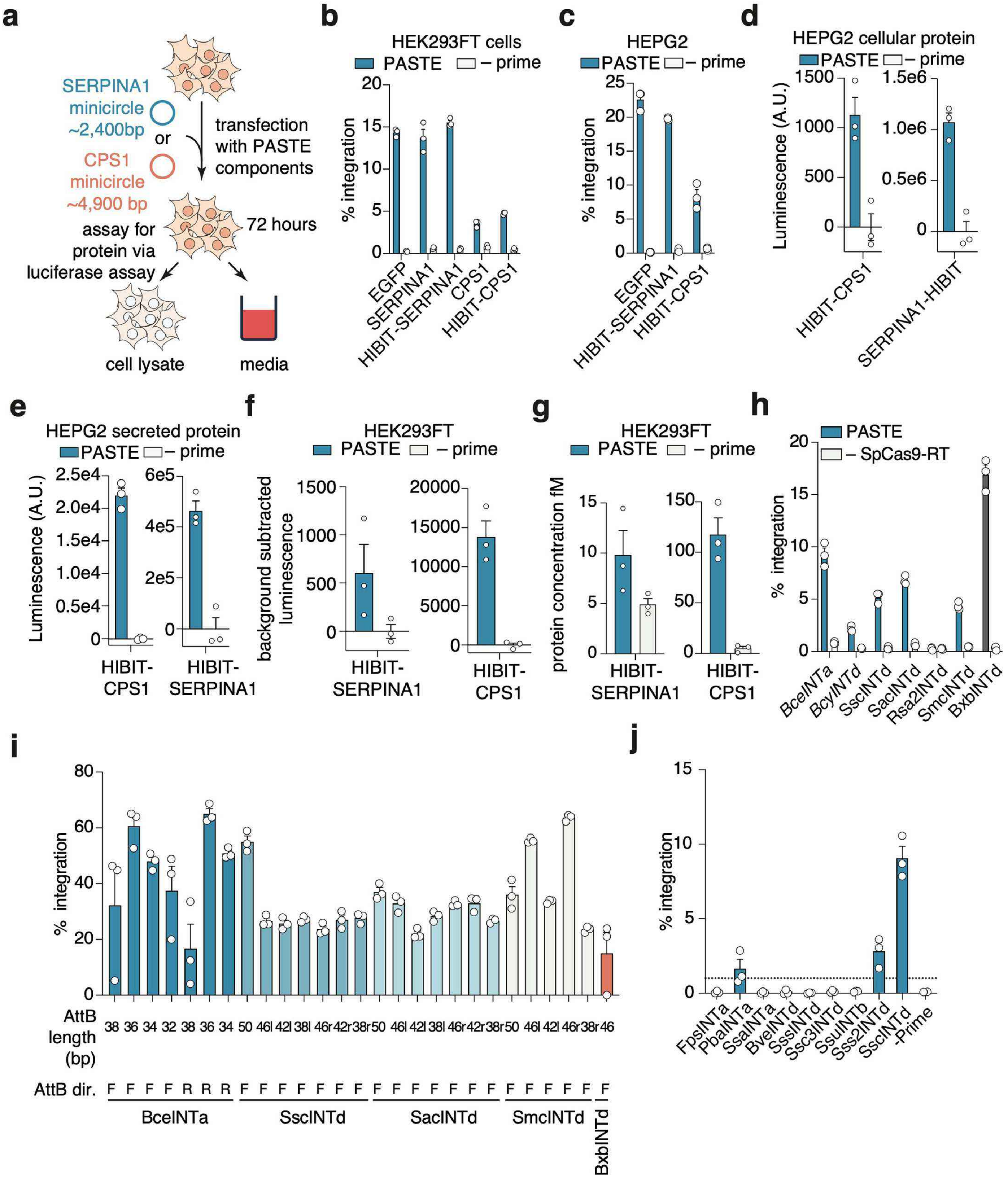
**Extended Data Fig. 6 | See next page for caption.**

**Extended Data Fig. 6 | Further characterization of PASTE specificity and effects on cellular transcriptome. a**) Comparison of indel rates generated by PASTE and HITI mediated insertion of EGFP at the *ACTB* and *LMNB1* loci in HepG2 cells. **b**) Effect of *attB* site integration on protein production. Samples treated with either *ACTB, LMNB1* non-targeting guides were harvest and analyzed for protein expression by Western blot. Quantified band intensities relative to *GAPDH* controls are shown below samples. **c**) GFP integration activity at predicted BxbINT and PASTE *ACTB* Cas9 guide off-target sites in the human genome. **d**) GFP integration activity at predicted HITI *ACTB* Cas9 guide off-target sites. **e**) Validation of ddPCR assays for detecting editing at predicted BxbINT off-target sites using synthetic amplicons. **f**) Validation of ddPCR assays for detecting editing at predicted PASTE *ACTB* Cas9 guide off-target sites using synthetic amplicons. **g**) Validation of ddPCR assays for detecting editing at predicted HITI *ACTB* Cas9 guide off-target sites using synthetic amplicons. **h**) Analysis of on-target and off-target integration events across 3 single-cell clones for PASTE and 3 single-cell clones for no prime condition. **i**) Volcano plots depicting the fold expression change of sequenced mRNAs versus significance (p-value). Each dot represents a unique mRNA transcript and significant transcripts are shaded according to either upregulation (red) or downregulation (blue). Fold expression change is measured against *ACTB*-targeting guide-only expression (including cargo). Significance is determined by moderated t-statistic[80] adjusted for a log-fold cut off of 0.585[81]. **j**) Top significantly upregulated and downregulated genes for BxbINT-only conditions. Genes are shown with their corresponding Z-scores of counts per million (cpm) for BxbINT only expression, GFP-only expression, PASTE targeting *ACTB* for EGFP insertion, Prime targeting *ACTB* for EGFP expression without BxbINT, and guide/cargo only. Data are mean (n = 3) ± s.e.m.



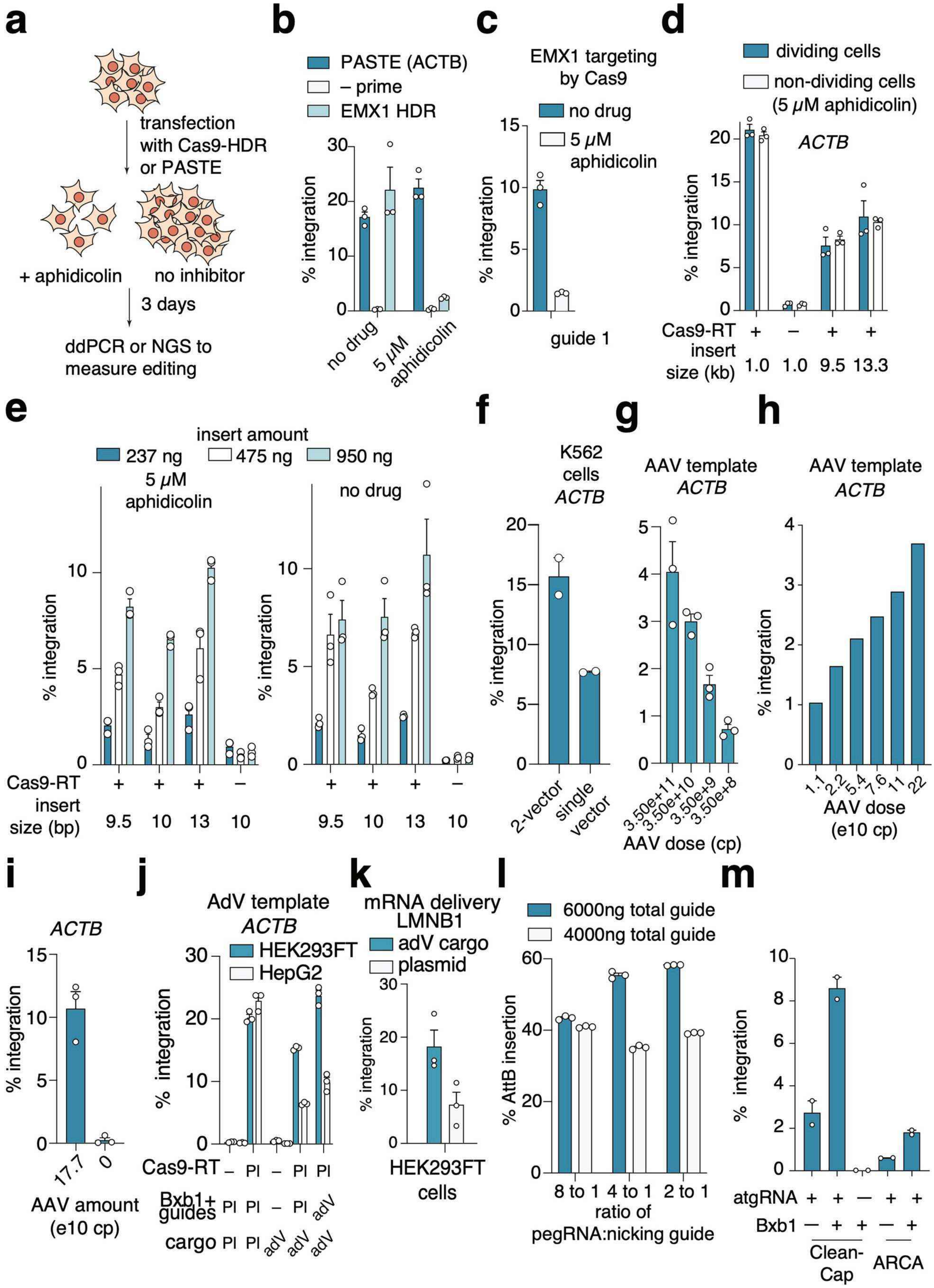
**Extended Data Fig. 7 | See next page for caption.**

**Extended Data Fig. 7 | Additional characterization of *attP* mutants for improved editing and multiplexing. a**) Integration efficiencies of wildtype and mutant *attP* sites with PASTE at the *ACTB* locus. **b**) *attP* single mutants are characterized for PASTE EGFP integration at the ACTB locus. **c**) Relative enrichment values (calculated as ratio of integrated reads to total reads) for the wildtype Bxb1 and top 5 mutants from the mutagenesis screen **d**) Comparison of integration efficiency between PASTEv3 and Twin-PE integration at the *ACTB* locus, with both single atgRNA (46 bp) or dual atgRNA with PASTE-Replace (38 bp). **e)** Comparison of integration efficiency and residual *attB* formation between PASTEv3 with PASTE-Replace and Twin-PE integration at the *NOLC1* locus with dual atgRNAs containing either a 46 bp or 42 bp *attB* sequence**. f)** Comparison of integration efficiency and residual *attB* formation between PASTEv3 with PASTE-Replace and Twin-PE integration at the *CCR5* locus with dual atgRNAs containing a 38 bp *attB* sequence. **g)** Comparison of residual *attB* formation between PASTEv3 with PASTE-Replace and Twin-PE integration at the *ACTB* locus. **h**) Characterization of integration of a 5 kb payload at the *ACTB* locus with all 16 possible dinucleotides for *attB*/*attP* pairs between the atgRNA and minicircle. **i)** Schematic of the pooled *attB*/*attP* dinucleotide orthogonality assay. Each *attB* dinucleotide sequence is co-transfected with a barcoded pool of all 16 *attP* dinucleotide sequences and BxbINT, and relative integration efficiencies are determined by next generation sequencing of barcodes. All 16 *attB* dinucleotides are profiled in an arrayed format with attP pools. **j)** Relative insertion preferences for all possible *attB*/*attP* dinucleotide pairs determined by the pooled orthogonality assay. **k)** Orthogonality of BxbINT dinucleotides as measured by a pooled reporter assay. Each web logo motif shows the relative integration of different *attP* sequences in a pool at a denoted *attB* sequence with the listed dinucleotide. **l**) Representative fluorescence images of multiplexed PASTE gene tagging of *ACTB, LMNB1*, and *NOLC1*. Data are mean (n = 3) ± s.e.m.



**Extended Data Fig. 8 | See next page for caption.**

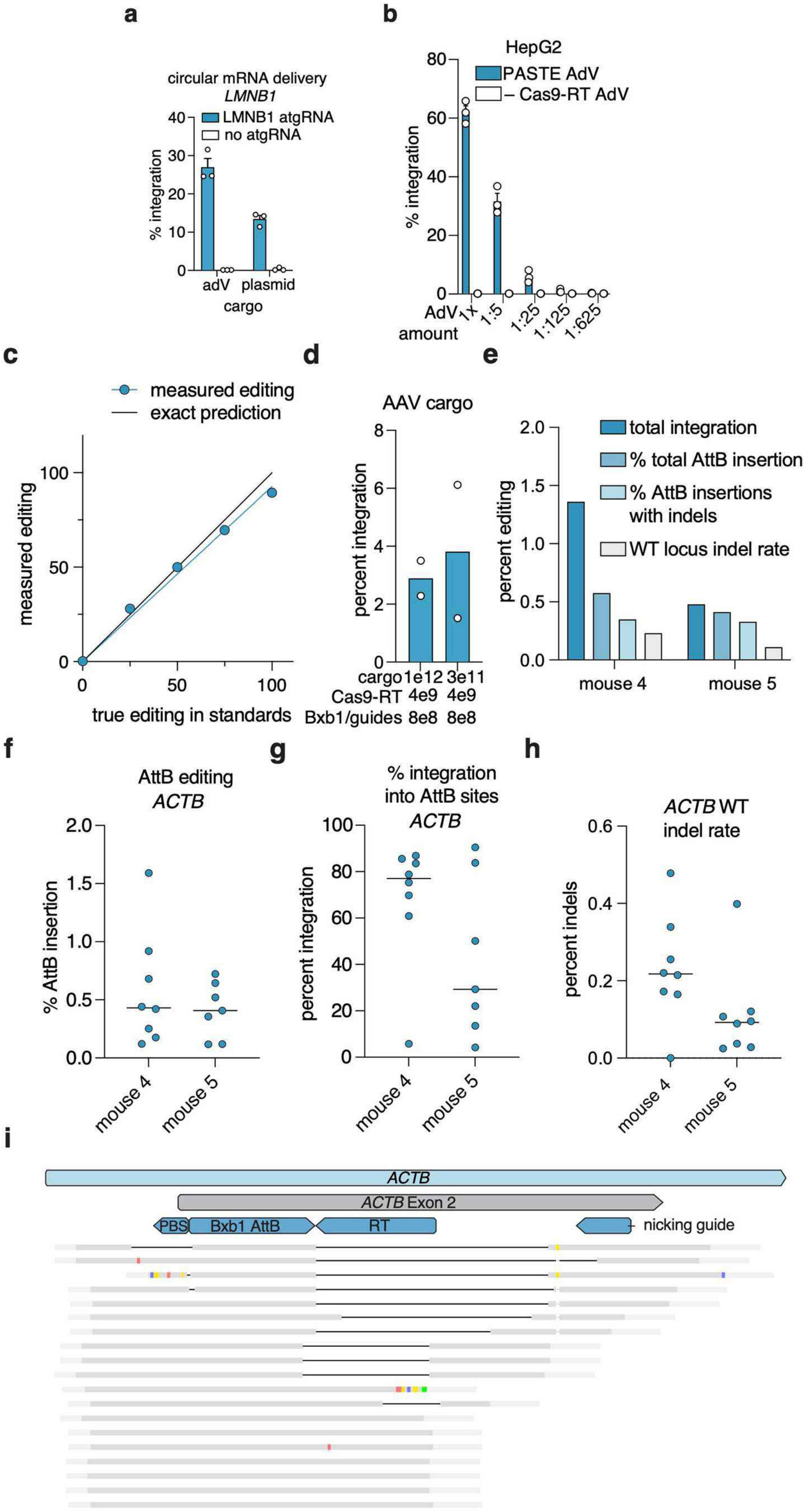
**Extended Data Fig. 8 | Therapeutic applications of PASTE and further characterization of integrases. a**) Schematic of protein production assay for PASTE-integrated transgene. *SERPINA1* and *CPS1* transgenes are tagged with HIBIT luciferase for readout with both ddPCR and luminescence. **b**) Integration efficiency of *SERPINA1* and CPS1 transgenes in HEK293FT cells at the *ACTB* locus. **c**) Integration efficiency of *SERPINA1* and CPS1 transgenes in HepG2 cells at the *ACTB* locus. **d**) Intracellular levels of SERPINA1-HIBIT and CPS1-HIBIT in HepG2 cells. **e**) Secreted levels of SERPINA1-HIBIT and CPS1-HIBIT in HepG2 cells. **f**) Integration of SERPINA1 and CPS1 genes that are HIBIT tagged as measured by a protein expression luciferase assay. **g)** Integration of SERPINA1 and CPS1 genes that are HIBIT tagged as measured by a protein expression luciferase assay normalized to a standardized HIBIT ladder, enabling accurate quantification of protein levels. **h**) PASTE integration activity with most active integrases compared to BxbINT. **i**) Characterization of integrase activity on truncated attachment sites using integrase reporters in HEK293FT cells**. j**) PASTE integration activity with computationally selected integrases with shorter *attB* sites. Data are mean (n = 3) ± s.e.m.



Extended Data Fig. 9 | See next page for caption.

**Extended Data Fig. 9 | Evaluation of viral templates for PASTE and characterization of editing in non-dividing cells. a**) Schematic of PASTE performance in the presence of cell cycle inhibition. Cells are transfected with plasmids for insertion with PASTE or Cas9-induced HDR and treated with aphidicolin to arrest cell division. Efficiency of PASTE and HDR are read out with ddPCR or amplicon sequencing, respectively. **b**) Editing efficiency of single mutations by HDR at EMX1 locus with two Cas9 guides in the presence or absence of cell division read out with amplicon sequencing. Data are mean (n = 3) ± s.e.m. **c**) HDR mediated editing of the EMX1 locus is significantly diminished in non-dividing HEK293FT cells blocked by 5 µM aphidicolin treatment. Data are mean (n = 3) ± s.e.m. **d**) Integration efficiency of various sized GFP inserts up to 13.3 kb at the *ACTB* locus with PASTE in the presence or absence of cell division. Data are mean (n = 3) ± s.e.m. **e)** Effect of insert minicircle DNA amount on PASTE-mediated insertion at the *ACTB* locus in dividing and non-dividing HEK293FT cells blocked by 5 µM aphidicolin treatment. Data are mean (n = 3) ± s.e.m. **f)** PASTE efficiency of EGFP integration at the *ACTB* locus in K562 cells. Data are mean (n = 3) ± s.e.m. **g**) Insertion templates delivered via AAV transduction. Templates were co-delivered via AAV dosing at levels indicated. Data are mean (n = 3) ± s.e.m. **h**) PASTE integration of GFP at the *ACTB* locus with the GFP template delivered via AAV in HEK293FT cells. **i)** PASTE integration of GFP at the *ACTB* locus with the GFP template delivered via AAV at different doses in HEK293FT cells. Data are mean (n = 3) ± s.e.m. **j)** Integration efficiency of AdV delivery of integrase, guides, and cargo in HEK293FT and HepG2 cells. BxbINT and guide RNAs or cargo were delivered either via plasmid transfection (Pl), AdV transduction (AdV), or omitted (-). SpCas9-RT was only delivered as plasmid or omitted. Data are mean (n = 3) ± s.e.m. **k**) Delivery of PASTE system components with mRNA and synthetic guides, paired with either AdV or plasmid cargo. Data are mean (n = 3) ± s.e.m. **l)** Attachment site insertion efficiency at the *LMNB1* locus using PASTE delivered as mRNA with synthetic atgRNA and nicking guides. Data are mean (n = 3) ± s.e.m. m) Integration efficiency at the *LMNB1* locus using PASTE delivered as mRNA (Trilink versions), synthetic atgRNA and nicking guides, and adenoviral delivered EGFP cargo. All conditions contain full length PASTE mRNA and are optionally supplemented with additional Bxb1 mRNA as indicated. Data are mean (n = 2) ± s.e.m.

https://doi.org/10.1038/s41587-022-01527-4



**Extended Data Fig. 10 | See next page for caption.**

https://doi.org/10.1038/s41587-022-01527-4

**Extended Data Fig. 10 | Additional characterization of *in vivo* liver editing with PASTE. a**) PASTE integration using delivery of circular mRNA with synthetic guides and either AdV or plasmid cargo. Data are mean (n = 3) ± s.e.m. **b**) PASTE integration of GFP at the *ACTB* locus with dose titration of PASTE components and GFP cargo delivered as AdV in HepG2 cells. Data are mean (n = 3) ± s.e.m. **c**) Evaluation of a 3-primer NGS assay for measuring integration efficiency, akin to junctional readouts by ddPCR. Using amplicon standards mixed at predefined ratios (x-axis), we can ascertain the accuracy of the measured editing (y-axis) by NGS. **d**) Analysis of primary human hepatocyte (PXB-cells®) EGFP integration at the *ACTB* locus using adenoviral delivery for PASTEv1 and guides and AAV for the EGFP template. Viral doses are as indicated. Shown is mean ± s.e.m with n = 2. **e**) Analysis of all liver editing outcomes for adenoviral EGFP template integration at the *ACTB* locus using PASTE *in vivo*. **f**) Analysis of *attB* site insertion efficiency at the *ACTB* locus using PASTE *in vivo*. Data are mean (n = 8). **g**) Analysis of adenoviral EGFP template integration efficiency into available *attB* sites at the *ACTB* locus using PASTE *in vivo*. Data are mean (n = 8). **h**) Analysis of indel frequency at the *ACTB* locus using PASTE *in vivo*. Data are mean (n = 8). **i**) Analysis of *attB*-site associated indels during *in vivo* integration with PASTE via alignment of representative reads to the *ACTB* locus containing the desired *attB* site.

nature portfolio

Corresponding author(s): Omar Abudayyeh, Jonathan Gootenberg

Last updated by author(s): Sep 10, 2022

# Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☒ | ☐ | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☒ | ☐ | A description of all covariates tested |
| ☒ | ☐ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☒ | ☐ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | Described in methods. |
| Data analysis | Described in methods. We used HMMER v3.3.2 in our analysis. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw reads for RNA sequencing and the atgRNA efficiency screen are available at Sequence Read Archive under BioProject accession number PRJNA700575. Expression plasmids are available from Addgene at https://www.addgene.org/browse/article/28223250/ under UBMTA. The human genome GRCh38 can be accessed at https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/.

# Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | We used a typical sample size for our field for cellular experiments, n=3. |
| Data exclusions | No data was excluded |
| Replication | We have repeated most experiments in the manuscript at least 3 times on different days. All attempts were successful. |
| Randomization | not applicable. There are no experiments that require randomization. |
| Blinding | not applicable. No experiments needed blinding for analysis. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study | |
|---|---|---|
| ☐ | ☒ | Antibodies |
| ☐ | ☒ | Eukaryotic cell lines |
| ☒ | ☐ | Palaeontology and archaeology |
| ☐ | ☒ | Animals and other organisms |
| ☒ | ☐ | Human research participants |
| ☒ | ☐ | Clinical data |
| ☒ | ☐ | Dual use research of concern |

## Methods

| n/a | Involved in the study | |
|---|---|---|
| ☒ | ☐ | ChIP-seq |
| ☒ | ☐ | Flow cytometry |
| ☒ | ☐ | MRI-based neuroimaging |

## Antibodies

| | |
|---|---|
| Antibodies used | For the primary stain, the primary antibodies were mixed with 1.25% goat serum and 300 μL were added per well according to the following dilutions: 1:1500 for the anti-ACTB antibody (NB600-501SS, NovusBio), 1:200 for the anti-SRRM2 antibody (NBP2-55697, NovusBio), 1:200 for the anti-NOLC1 antibody (11815-1-AP, ProteinTech), and 1:200 for the anti-LMNB1 antibody (12987-1-AP, ProteinTech). After shaking overnight at 4 °C, the wells were washed three times with 1mL PBS pH 7.4. For the secondary staining, 1:1000 dilution of secondary antibody, either goat anti-mouse IgG Alexa Fluor 568 (Thermo Fisher Scientific, A-11004) or goat anti-rabbit IgG Alexa Fluor 647 (Thermo Fisher Scientific, A21244), were mixed with 1.25% goat serum.<br>For western blots: Membranes were then incubated with primary antibodies (β-Actin antibody [4967S] with GAPDH antibody [97166S] or Lamin B1 antibody [12586S] with GAPDH antibody [97166S]) from Cell Signaling Technology. Membranes were washed for four 5-min washes with PBS-T (0.2% Tween-20) and further incubated with LICOR Donkey Anti-Rabbit IgG Polyclonal Antibody (IRDye® 800CW) and Goat Anti-Mouse IgG Polyclonal Antibody (IRDye® 680RD) diluted 1:15,000 in Intercept (PBS) Blocking Buffer with 0.2% Tween-20. |
| Validation | All primary antibodies were validated by the manufacturer. Please see citations and data on each of the product pages using the product codes listed in the previous section. |

## Eukaryotic cell lines

Policy information about cell lines

| | |
|---|---|
| Cell line source(s) | Cell lines were retrieved Thermo Fisher (HEk239FT) or ATCC (HepG2, K562). Primary hepatocytes were from Thermo Fisher. Primary T cells were from Stemcell Technologies |
| Authentication | None of the cell lines were authenticated by us. We bought them commercially. |

Mycoplasma contamination

Cell lines were not tested for mycoplasma.

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified lines were used.

## Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

The FRG KO model possesses a triple knockout of the Fah, Rag2 and Il2rg genes. The FRG genotype enables the animals to be engrafted and repopulated with human hepatocytes. For this study, we used female liver humanized Fah-/-/Rag2-/-/Il2rg-/- KO on C587BL/6 with ≥70% human hepatocyte repopulation. Mice were 5.5 months at time of injection.

The FRG mice were maintained in an environment that was designed to support their well being. Monitoring systems confirmed that the vivarium was within acceptable temperature, pressure, and relative humidity tolerances.

Since the FRG mice maintained in the study were immuno-compromised, they were housed within a purified air environment designed explicitly to maintain their health and well-being. Entry into animal room, vivarium, was gained through an antechamber, which limited any air flowing directly between the laboratory from the HEPA purified air of the vivarium.

There were multiple stages or redundancies to the purified air environment during the study:
Stage 1: The caging rack systems provided HEPA purified air via an integrated air handling fan and filtration system. Purified air passed through the cage at approximately 65 ACH, and the cage rack system drew from the air within a portable HEPA environment that makes up Stage 2.
Stage 2: HEPA purified air was provided by a portable HEPA environment known as a "BioBubble." This enclosure provided 75-100 air changes per hour during the study.
Stage 3: The HVAC system also provided HEPA filtration of all air introduced into the vivarium outside the BioBubble enclosure.

Humidity was monitored and controlled through the HVAC system to maintain relative humidity levels at comfortable levels for the mice, between 30-70%.

Temperature: The HVAC system maintained ambient temperatures of 22°C +/- 8°C during the study.

Illumination: The light and dark phases provided a regular diurnal cycle for the animals. A timing device controlled the cycling of light within the vivarium and the lights cycled at 12-hour intervals, from 7 AM to 7 PM. Illumination within the vivarium ranged from 45-30 ft candles within the BioBubble environment.

Wild animals

No wild animals were used in this study.

Field-collected samples

no field collected samples were used in the study.

Ethics oversight

Mice were maintained at Yecuris Corporation affiliated IACUC accredited facility. General procedures for animal care and housing were as described in the Guide for the Care and Use of Laboratory Animals, National Research Council, Yecuris IACUC Policy and Yecuris General Mouse Handling Care and Euthanasia. Cages were changed every two weeks and the testing facility was sanitized weekly. Animal studies were carried out in accordance with the recommendations in the Guide for the Care and Use of Laboraty Animals of the National Institutes of Health. The protocols were approved by the Institutional Animal Care and Use Committee at the Massachusetts Institute of Technology (Protocol Number: 0919-065-22) and Yecuris Corporation IACUC Policy.

Note that full information on the approval of the study protocol must also be provided in the manuscript.