A versatile platform for locus-scale genome rewriting and verification

Ran Brosh[a,1], Jon M. Laurent[a,1,2], Raquel Ordoñez[a], Emily Huang[a], Megan S. Hogan[a], Angela M. Hitchcock[a], Leslie A. Mitchell[a,3], Sudarshan Pinglay[a], John A. Cadley[a], Raven D. Luther[a], David M. Truong[a,3], Jef D. Boeke[a,b,c,4], and Matthew T. Maurano[a,d]

[a]Institute for Systems Genetics, NYU Langone Health, New York, NY 10016; [b]Department of Biochemistry and Molecular Pharmacology, NYU Langone Health, New York, NY 10016; [c]Department of Biomedical Engineering, NYU Tandon School of Engineering, Brooklyn 11201, NY; and [d]Department of Pathology, NYU Langone Health, New York, NY 10016

Contributed by Jef D. Boeke, January 14, 2021 (sent for review November 19, 2020; reviewed by Barak A. Cohen and Evgeny Kvon)

Routine rewriting of loci associated with human traits and diseases would facilitate their functional analysis. However, existing DNA integration approaches are limited in terms of scalability and portability across genomic loci and cellular contexts. We describe Big-IN, a versatile platform for targeted integration of large DNAs into mammalian cells. CRISPR/Cas9-mediated targeting of a landing pad enables subsequent recombinase-mediated delivery of variant payloads and efficient positive/negative selection for correct clones in mammalian stem cells. We demonstrate integration of constructs up to 143 kb, and an approach for one-step scarless delivery. We developed a staged pipeline combining PCR genotyping and targeted capture sequencing for economical and comprehensive verification of engineered stem cells. Our approach should enable combinatorial interrogation of genomic functional elements and systematic locus-scale analysis of genome function.

genome engineering | genome writing | regulatory genomics | stem cells | synthetic biology

A global understanding of genomic regulatory architecture is critical to interpreting the effect of variants associated with common human traits and diseases (1). As the regulation of genes throughout development depends strongly on their native chromatin and genomic environments (2), short artificial constructs are inherently incapable of modeling the complexity of native loci, even when integrated genomically. Analysis of natural sequence variation in regulatory DNA provides one high-throughput approach for functional assessment in an endogenous cellular and genomic context, but detailed investigation of locus architecture is limited by the low frequency of informative variants and patterns of linkage disequilibrium (3, 4).

Transgenic mammalian cell lines and animals generated using homologous recombination (5, 6) and the subsequent development of nuclease-mediated genome editing (7) have enabled functional analysis of the regulation of model genes at their endogenous loci. Extensions of these technologies have since facilitated screens of noncoding regulatory elements (8, 9) and locus-scale analyses (10, 11). However, editing approaches offer limited control over the final sequence, a low maximum edit size limited by the difficulty of multiplexed editing, no inherent allele specificity at diploid loci, and the risk of off-target editing by designer nucleases (12). Many limitations of genome editing do not apply to production of DNA using recombineering or yeast assembly approaches (13, 14). Indeed, transgenesis of large constructs, such as yeast and bacterial artificial chromosomes (YACs and BACs) (15), has enabled position-independent, copy-number–dependent expression, reproduction of organismal phenotypes, such as the developmental switch from fetal to adult hemoglobin (16, 17), and modeling of disease-associated variation (18).

Recombinase-mediated cassette exchange (RMCE) (19–22) and serine recombinase approaches (23) have enabled efficient single-copy targeting in mammalian cells, and have been adapted for delivery of large DNAs (24, 25). But while gene function is tightly linked to cellular and genomic context, existing delivery schemes are not readily portable to new loci or cell lines. Human and mouse embryonic stem cells (hESCs/mESCs) offer a scalable platform for assessment of gene function, as they can be differentiated in vitro to a wide variety of cell types, and mESCs rapidly yield transgenic animals through tetraploid complementation (26, 27). However, rapid engineering is impeded by their stringent growth requirements and intolerance of certain selection schemes. Furthermore, existing approaches do not address the challenge of verifying both on-target and off-target events. Finally, the gene traps employed in some RMCE schemes to select for integrants remain as transcriptionally active genomic scars, which may impact the function of nearby regulatory elements. Overcoming these obstacles to rewriting endogenous loci would permit a synthetic approach to regulatory genomics, where transgenic analysis facilitates dissection of the regulatory architecture of mammalian genomes.

Significance

Functional analysis of noncoding genomic regulatory elements, which harbor the majority of common human disease and trait associations, is complicated by their cellular and genomic context sensitivity. We developed Big-IN, a method for rewriting large segments of mammalian genomes, including full genes and their surrounding regulatory elements. We demonstrate a flexible genomic verification pipeline to identify correctly engineered cells. We expect Big-IN will enable technologies for synthesis and assembly of large DNAs to catalyze a synthetic approach to regulatory genomics.

Author contributions: R.B., J.M.L., S.P., J.D.B., and M.T.M. designed research; R.B., J.M.L., R.O., E.H., M.S.H., A.M.H., L.A.M., and R.D.L. performed research; R.B., J.M.L., R.O., L.A.M., J.A.C., D.M.T., and J.D.B. contributed new reagents/analytic tools; R.B., J.M.L., R.O., J.A.C., J.D.B., and M.T.M. analyzed data; and R.B., J.D.B., and M.T.M. wrote the paper.

Reviewers: B.A.C., Washington University in St. Louis School of Medicine; and E.K., University of California, Irvine.

Competing interest statement: R.B., J.M.L., J.D.B., and M.T.M. are listed as inventors on a patent application describing Big-IN. L.A.M. is a founder of Neochromosome, Inc. and is on the Scientific Advisory Board of ReOpen Diagnostics. D.M.T. is an employee of Neochromosome, Inc. J.D.B. is a founder and director of CDI Labs, Inc., a founder of Neochromosome, Inc., a founder of and consultant to ReOpen Diagnostics, and serves or served on the Scientific Advisory Board of Sangamo, Inc., Modern Meadow, Inc., and Sample6, Inc.

[1]R.B. and J.M.L. contributed equally to this work.

[2]Present address: Department of Research and Development, Pandemic Response Lab NYC, New York, NY 10016.

[3]Present address: Department of Research and Development, Neochromosome, New York, NY 10016.

[4]To whom correspondence may be addressed. Email: jef.boeke@nyulangone.org.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2023952118/-/DCSupplemental.

Published March 1, 2021.

Here we describe Big-IN, a modular RMCE platform for synthetic regulatory genomics. We developed a pipeline for efficient engineering of hESCs and mESCs with a Big-IN landing pad (LP) at a target locus. This LP facilitates single-step allelic integration and enables repeated delivery to the same locus. We further describe a scalable validation pipeline based on PCR genotyping and targeted sequencing for unbiased verification of the engineered cells prior to functional analysis.

Results

Engineering the *HPRT1* Locus in Human ESCs. To enable repeated, precise, and efficient delivery of large DNAs to a given locus, we employed a two-stage approach that first targets a short LP to replace a genomic locus of interest using CRISPR/Cas9-mediated homology directed repair (HDR) (Fig. 1*A*). A plasmid (pLP-TK) was engineered to include the human EF1α promoter (pEF1α) to drive ubiquitous expression of a single open reading frame (ORF) comprising a puromycin-resistance gene (PuroR) fused to a truncated Herpes simplex virus thymidine kinase (HSV1-ΔTK) gene (28) and a Cre^{ERT2} gene (29), separated by a P2A peptide (30). Interposed between the LP ORF and the vector backbone are heterotypic loxM (lox 2272) and loxP sites to permit subsequent RMCE. The lox sites are flanked by homology arms (HAs) corresponding to the genomic sequences flanking guide RNA (gRNA) target sites at the targeted genomic locus. To facilitate clearance of the transiently transfected plasmid by inducing its linearization in vivo, the same gRNA target sequences and protospacer adjacent motifs (PAM) were cloned into the vector backbone just outside the HAs.

We targeted the X-linked *HPRT1* locus for LP integration to permit counterselection with the cytotoxic antimetabolite 6-Thioguanine (6-TG) (31). H1 male hESCs, which harbor a single copy of *HPRT1*, were cotransfected with pLP-TK and pCas9 plasmids (32) expressing gRNAs targeting a 42-kb region, including the *HPRT1* gene for replacement. Cells were sequentially treated with 6-TG and puromycin to select for *HPRT1* loss and LP-TK gain, followed by clonal isolation. Correct LP-TK integration was verified by PCR genotyping using primers targeting the novel junctions between LP-TK and the genomic sequences beyond the HAs (Fig. 1*B*). A candidate clone (58I) was selected for further validation. Junction PCR amplicons were subjected to Sanger sequencing, to verify correct LP-TK integration at base pair resolution (*SI Appendix*, Fig. S1*A*). Quantitative real-time PCR (qRT-PCR) confirmed loss of *HPRT1* mRNA expression and gain of Cre^{ERT2} expression (*SI Appendix*, Fig. S1*B*). Robust cytotoxic activity of HSV1-ΔTK following ganciclovir (GCV) treatment was validated in a kill curve (*SI Appendix*, Fig. S1*C*). We also developed a lentiviral reporter assay for Cre activity, which indicated that Cre^{ERT2} is rapidly and efficiently activated by tamoxifen (*SI Appendix*, Fig. S1*D*). Thus, the function of all three components of the LP ORF was verified.

To facilitate comprehensive genomic verification of multistep cellular engineering with these complex constructs, we developed a modular next-generation sequencing analysis approach, which independently maps short reads to both reference genomes (hg38 and mm10) and custom references for each engineering construct. We further applied hybridization capture sequencing (Capture-seq) approach to efficiently verify correct engineering of screened clones (Fig. 1*C*). We employed nick translation to generate bait in a rapid, flexible, and cost-effective fashion. Using this mapping pipeline, whole-genome sequencing (WGS) of clone 58I verified loss of the targeted *HPRT1* locus, gain of LP-TK, and absence of LP-TK backbone and pCas9 (Fig. 1 *D*–*F*).

Integration relied on 1-kb HAs to correctly target the LP, but HA length reduces the efficiency of PCR genotyping from genomic DNA (Fig. 1*B*) and impedes the mapping of short sequencing reads that definitively span the LP-HA/genome junctions. Therefore, we measured relative integration efficiency with shorter HAs. We integrated a series of pLP-TK plasmids with varying HA lengths and estimated on-target integration as the relative number of cells surviving puromycin and 6-TG selection, revealing that efficient integration could be performed with HAs as short as 100 bp (*SI Appendix*, Fig. S1*E*), facilitating subsequent sequence-based mapping of integration sites.

We also assessed the efficacy of our in vivo linearization strategy to reduce off-target integration of transiently transfected plasmids. We designed two pLP-TK plasmids differing only in the presence of the LP-flanking gRNA sites required for in vivo linearization, targeted them to *HPRT1*, selected for correct integrants with puromycin and 6-TG, and subjected the pool of cells to Capture-seq. We found that the relative coverage depth of the LP backbone was lower for the in vivo-linearized pLP-TK (*SI Appendix*, Fig. S1*F*), possibly due to enhanced HDR efficiency (33) and reduced plasmid half-life (which was evident from shortened transient puromycin resistance of the transfected cells).

Delivery of large DNA through cassette exchange is an infrequent event, requiring selection to obtain practical efficiency. The HSV1-ΔTK gene encoded by LP-TK is a widely used counterselectable marker that renders cells sensitive to GCV by converting it to the toxic metabolite GCV-triphosphate (GCV-TP), which inhibits DNA synthesis and leads to cell death (34). To demonstrate a counterselection-based approach to isolation of successful RMCE events, we designed a minimal 2.7-kb payload (PL1), comprising a pEF1α-driven GFP-T2A-BSD (blasticidin S deaminase) ORF flanked by loxM and loxP sites (Fig. 1*G*). H1 LP-TK cells were transfected with a PL1-harboring plasmid (pPL1) and LP-derived Cre^{ERT2} activity was induced with tamoxifen. Cells were selected with blasticidin to enrich for PL1-expressing cells, followed by GCV counterselection of TK-expressing cells. PCR genotyping of isolated clones showed a 100% rate of replacement of LP-TK with PL1 (Fig. 1*H*). Capture-seq analysis of four selected clones confirmed the presence of PL1, the absence of any plasmid backbone, and the loss of LP-TK (Fig. 1 *I* and *J*). The integrated PL1 was transcriptionally active, as evident from GFP expression (*SI Appendix*, Fig. S1*G*).

Efficient Counterselection for Delivery. To quantify the efficacy of TK/GCV counterselection in H1 hESCs, we mixed TK^- and TK^+ (LP-TK) cells at different ratios and treated these cocultures with GCV. More than 80% of the TK^- cells died when mixed at a 1:1 ratio with TK^+ cells, and all died when mixed at a 1:10 ratio (Fig. 2*A*). Indeed, it is known that GCV-TP can diffuse from TK^+ cells to TK^- cells via gap junctions (35, 36). The resulting bystander cell death in TK^- cells limits the ability to recover rare events (Fig. 2*B*).

Therefore, we tested an alternative counterselection strategy (Fig. 2*C*) that relies on the X-linked *PIGA* (phosphatidylinositol glycan anchor biosynthesis class A) gene, which encodes an enzyme crucial for the biosynthesis of glycosylphosphatidylinositol (GPI) anchors (37) and renders cells sensitive to proaerolysin, a bacterial prototoxin. Proaerolysin perforates the plasma membrane upon binding to GPI anchors on the cell surface, resulting in rapid cell death (38). Furthermore, PIGA activity can be quantitatively monitored by measuring levels of CD59, a broadly expressed membrane-linked GPI-anchored protein (39). Deletion of *PIGA* can be selected for with proaerolysin after a short period to allow for loss of PIGA protein and subsequent loss of GPI-anchored proteins from the cell surface (40).

While proaerolysin efficiently killed parental H1 hESCs, Δ*PIGA* cells, in which the *PIGA* gene was deleted using CRISPR/Cas9 (Materials and Methods), were entirely resistant (*SI Appendix*, Fig. S2). Integration of an LP expressing a human mini *PIGA* gene (hmPIGA) to the *HPRT1* locus resensitized H1 Δ*PIGA* hESCs to proaerolysin and restored CD59 expression (*SI Appendix*, Fig. S2 *B* and *F*). Importantly, rare Δ*PIGA* H1 hESCs were efficiently isolated when cocultured with parental H1 cells by applying proaerolysin selection (Fig. 2*D*). This suggested that

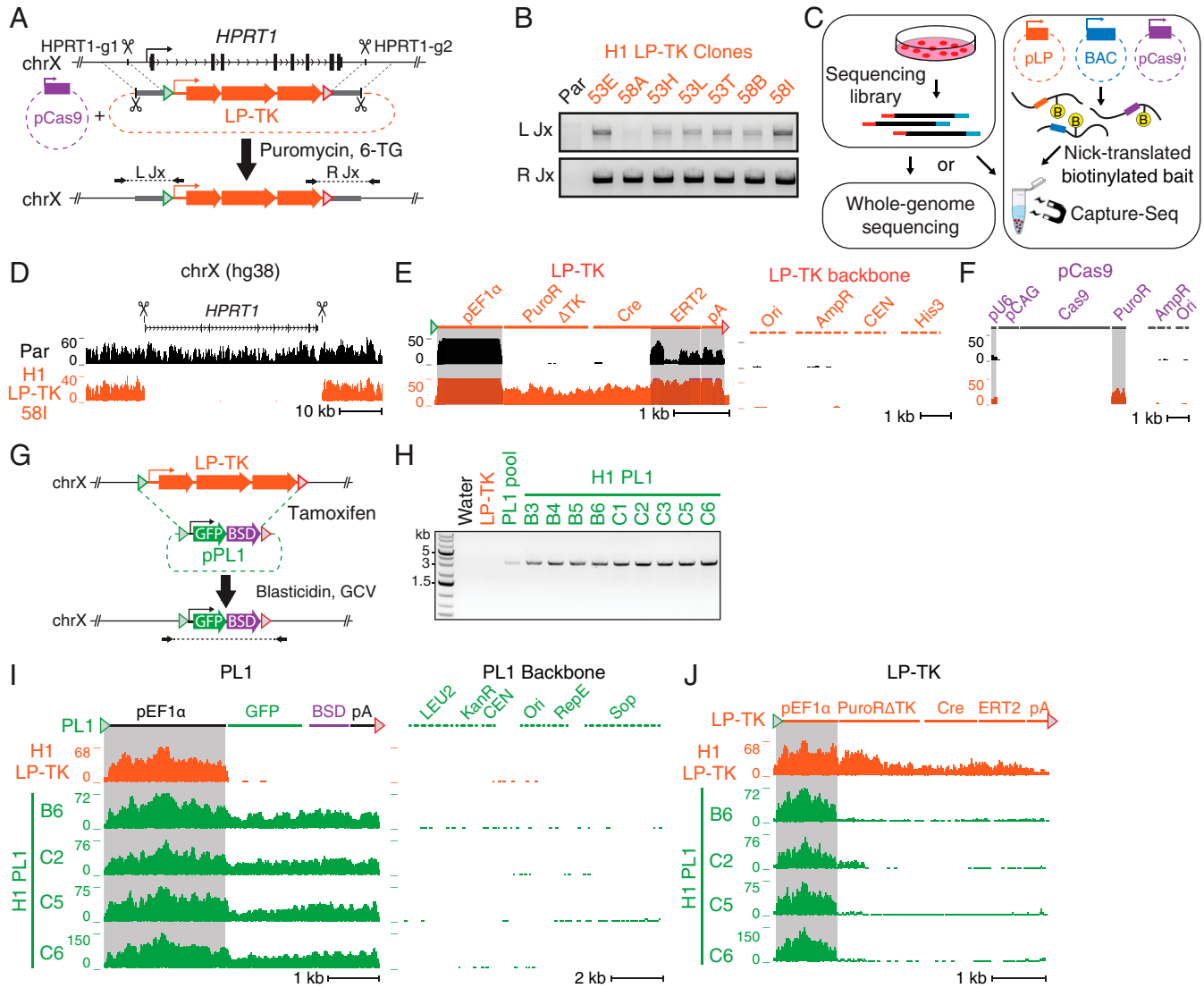

Fig. 1. Engineering the *HPRT1* locus in hESCs. (*A*) Replacement of the 42-kb *HPRT1* locus in H1 hESCs with an LP (LP-TK) utilizing CRISPR/Cas9 and 1-kb HAs (gray). Cells are selected for LP-TK presence with puromycin and HPRT1 inactivation with 6-TG. (*B*) PCR genotyping of H1 clones for novel left (L) and right (R) junctions (Jx) using primers illustrated in *A*. Par, parental H1. (*C*) Sequencing verification pipeline using WGS or targeted libraries. Capture-seq enriches for regions of interest using biotinylated bait prepared using nick translation from relevant DNA constructs. (*D*) WGS of parental H1 hESCs and LP-TK clone 58I mapped to hg38 shows the 42-kb deletion of the *HPRT1* locus. (*E*) Mapping to LP-TK (*Left*) and LP-TK backbone (*Right*) confirms specific gain of LP-TK; regions cross-mapping with human genome are shaded gray [pEF1α, *EEF1A1* promoter; ERT2, *ESR1* ligand binding domain (59); pA, *EIF1* pA signal]. (*F*) Mapping to pCas9 confirms plasmid loss; regions shaded gray cross-map with human (pU6, *U6* promoter) and LP-TK (PuroR, puromycin-resistance gene). (*G*) LP-TK at *HPRT1* undergoes RMCE with PL1 following transfection and Cre induction. Payload integration can be selected for with blasticidin and GCV. (*H*) Genotyping of untransfected LP-TK hESCs (clone 58I), PL1-transfected pool, and PL1 clones using PCR primers flanking payload lox sites (illustrated in *G*). All clones produced the expected 3-kb product (a 5.7-kb product for LP-TK cells was not detected at this extension time). (*I*) Capture-seq analysis of chosen H1 PL1 clones mapped to PL1 (*Left*) and its backbone (*Right*). (*J*) Capture-seq reads mapped to LP-TK, validating LP loss in PL1 clones. Cross-mapping sequences are shaded gray.

LP-expressed hmPIGA permits negative selection of LP-PIGA cells to effectively enrich for correct delivery events.

Recovery of rare events where a payload replaces the LP requires that expression of hmPIGA is stably maintained following withdrawal of positive selection. However, while nearly all H1 LP-PIGA cells maintained high CD59 levels in the presence of puromycin, a substantial proportion of cells spontaneously lost CD59 following puromycin withdrawal (*SI Appendix*, Fig. S2*F*) and showed reduced LP transcriptional activity (*SI Appendix*, Fig. S2*G*). Thus, any counterselection-based delivery and screening scheme must address a potentially high background of false-positive cells from LP silencing.

Allele-specific Engineering of the Murine *Sox2* Locus. To develop an approach for allele-specific engineering of diploid loci, we employed C57BL6/J × CAST/EiJ (BL6xCAST or BL6xC) F1 hybrid mESC cells (41), the genome of which harbors heterozygous point variants every 140 bp on average (42). We targeted the *Sox2* locus, which encodes a master transcription factor essential for regulation of pluripotency and differentiation (43, 44). We designed gRNAs targeting the flanks of a 143-kb genomic region that includes the *Sox2* coding sequence, promoter, long-distance regulatory regions, and several noncoding genes (44, 45). These gRNAs target BL6-specific PAMs to facilitate allele-specific engineering. We constructed pLP-PIGA to support

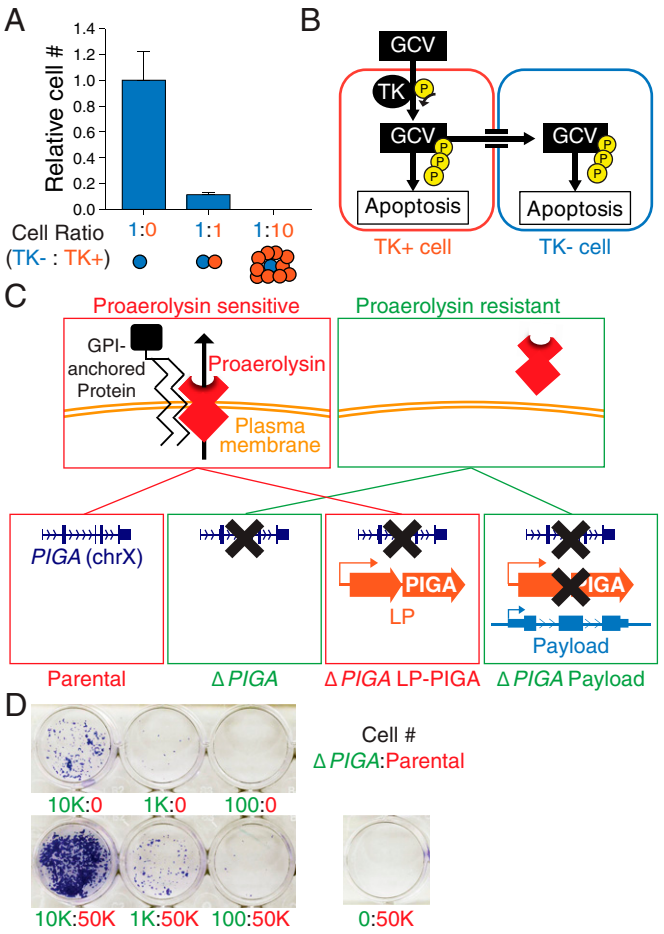

Fig. 2. Development of an efficient counterselection strategy. (*A*) Parental (TK^-) and LP-TK (TK^+) H1 hESCs were cocultured at the indicated ratios, treated with 1 μM GCV for 4 d, and assayed for the number of live cells using PrestoBlue. Cell counts are shown relative to unmixed parental cells. Bars show mean + SD ($n = 2$). (*B*) GCV enters TK^+ cells and is metabolized into the toxic membrane-impermeable compound GCV-TP, which diffuses into neighboring cells and induces bystander cell death in TK^- cells. (*C*) Big-IN counterselection strategy using PIGA/proaerolysin. (*D*) Parental and Δ*PIGA* H1 hESCs cocultured at the indicated ratios for 3 d were treated with 1 nM proaerolysin for 1 d and stained with Crystal violet 3 d later.

counterselection-based delivery to cell lines lacking a functional *Piga* gene (Fig. 2*C*). The LP ORF includes four components, each separated by three mutually recoded P2A peptides: mScarlet (46), Cre^{ERT2}, PuroR, and hmPIGA (Fig. 3*A*). The ORF is flanked by heterotypic loxM/loxP sites, short HAs, and gRNA target sites.

We transfected pLP-PIGA and pCas9 plasmids into BL6xCAST *ΔPiga* mESCs, selected cells with puromycin, and isolated clones. Of 40 clones screened using PCR genotyping, 16 (40%) contained both novel junctions (Fig. 3*B*). Passing clones were further screened with primers to detect Ori (common to multiple vector backbones), which eliminated eight Ori^+ clones (50%), likely resulting from retention or off-target integration of LP-PIGA backbone or pCas9. We confirmed the allele-specific loss of *Sox2* in 15 (94%) of the 16 clones using a BL6 allele-specific primer harboring 4 mismatched base pairs relative to the CAST allele (*SI Appendix*, Table S1).

A successful LP-PIGA integration (clone A1) and a clone that failed PCR genotyping were subjected to Capture-seq using bait generated from a BAC covering the *Sox2* region, and the pLP-PIGA and pCas9 plasmids. Inspection of coverage depth at the 143-kb *Sox2* genomic locus revealed a 50% reduction for clone A1 compared with parental mESCs or the failed clones (Fig. 3*C*), as expected for complete loss of the targeted BL6 allele. Clone A1 also showed specific gain of LP-PIGA with no coverage of the LP-PIGA backbone or pCas9, whereas the failed clone showed clear presence of the LP-PIGA backbone (Fig. 3 *D* and *E*). Expression of LP-PIGA components and BL6 allele-specific loss of *Sox2* expression in clone A1 was verified through qRT-PCR analysis (*SI Appendix*, Fig. S3*A*), which was chosen for future payload deliveries. We confirmed efficient isolation of rare mESCs using the *Piga*/proaerolysin counterselection strategy (*SI Appendix*, Fig. S3 *B* and *C*), and observed a similar silencing effect to LP-TK in the absence of positive selection (*SI Appendix*, Fig. S3*D*). In summary, we have demonstrated an efficient strategy for allele-specific LP integration and a comprehensive pipeline for verification of correctly engineered cells.

Efficient Delivery to mESCs. We attempted to deliver payloads to LP-PIGA mESCs using a positive/negative selection strategy. However, all clones that survived blasticidin and proaerolysin selection manifested multiple-copy payload gain, including its vector backbone, and without LP-PIGA loss (*SI Appendix*, Fig. S4*A*). We transiently augmented Cre activity through cotransfection of a Cre expression plasmid (pCAG-Cre). Additionally, we cloned a ΔTK expression cassette (BBTK) into the payload backbone to permit GCV-based counterselection against surviving colonies harboring off-target integrants. Cotransfection of pPL1-BBTK and pCAG-Cre readily resulted in efficient PL1 integration (Fig. 4*B*). To assess efficiency of larger payloads, $pSox2^{46kb}$-MC-BBTK was constructed including a 46-kb region of the *Sox2* locus and containing a marker cassette to enable positive selection (Fig. 4*A*). Upon delivery and selection, PCR genotyping verified that 99% of clones harbored correct payload integration (Fig. 4*B*). Six PCR-validated clones of each payload type were then chosen for Capture-seq analysis. Mapping sequencing reads to the PL1 sequence or mouse genome revealed that all clones had complete coverage of the delivered payload (Fig. 4*C*). In $Sox2^{46kb}$-MC clones, coverage depth was restored to parental levels over the genomic region corresponding to $Sox2^{46kb}$, while the remaining 97 kb of the *Sox2* deletion was unaffected (Fig. 4*D* and *SI Appendix*, Fig. S4*B*). Analysis of known CAST single nucleotide variants (SNVs) further confirmed reintroduction of BL6 alleles. There was no evidence for the gain of the payload backbone in any of the clones analyzed (*SI Appendix*, Fig. S4*C*), and all 79 clones lost LP-PIGA (*SI Appendix*, Fig. S4*D*). Selected PL1 and $Sox2^{46kb}$-MC cells both expressed the payload-derived BSD, while $Sox2^{46kb}$-MC clones also partially restored the expression of the BL6 allele of *Sox2* (*SI Appendix*, Fig. S4*E*). In addition, both cell types showed expression of payload-derived GFP (*SI Appendix*, Fig. S4*F*).

This approach leaves a BSD-GFP transcriptional unit (TU) integrated with the payload, which might affect the activity of nearby genes or regulatory elements. To develop an alternate architecture and selection strategy for scarless delivery, we constructed $pSox2^{143kb}$, which harbors the entire 143-kb *Sox2* BL6 allele replaced by LP-PIGA, and in which the BSD-GFP TU is relocated on the vector backbone, outside the lox sites (Fig. 4*A*). We delivered $pSox2^{143kb}$ to LP-PIGA mESCs together with pCAG-iCre, which encodes a codon-optimized Cre recombinase, and selected cells transiently with blasticidin to enrich for payload-transfected cells, followed by proaerolysin selection to eliminate unrecombined LP-PIGA mESCs. PCR genotyping identified four clones that lost LP-PIGA, one of which (G11) was positive for the newly formed BL6 allele genomic junctions (Fig. 4*E*). Capture-seq analysis verified the restoration of the entire 143-kb BL6 allele in clone G11, without gain of the payload backbone (Fig. 4*F*). Finally, qRT-PCR analysis confirmed

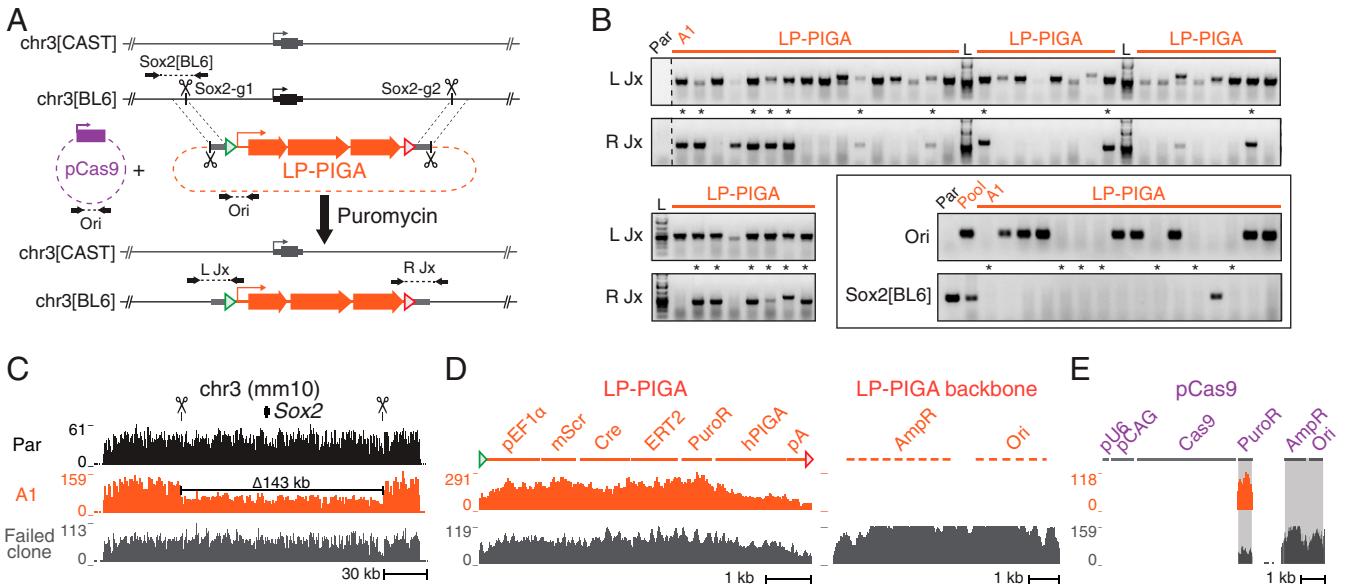

Fig. 3. Allele-specific engineering of the murine Sox2 locus. (*A*) Replacement of a 143-kb region of the *Sox2* locus on the BL6 allele of chromosome 3 (black) in BL6xCAST Δ*Piga* mESCs with of LP-PIGA utilizing CRISPR/Cas9, facilitated by 0.15-kb HAs (gray). (*B*, *Upper* and *Lower Left*) Screening of BL6xCAST LP-PIGA clones using PCR genotyping primers targeting novel junctions (L Jx and R Jx) illustrated in *A*. (*Lower Right*) Secondary screening of 16 clones positive for both junctions using primers for plasmid origin of replication (Ori) and the BL6 *Sox2* allele (Sox2[BL6]). Seven clones marked with asterisks had the desired genotype; Clone A1 was selected for further analysis. L, ladder; Par, parental BL6xCAST mESCs. (*C–E*) Capture-seq analysis of parental Δ*Piga* BL6xCAST mESCs, LP-PIGA clone A1, and an example failed clone from an independent LP-PIGA delivery. Reads were mapped to the references indicated above. Cross-mapping sequences are shaded gray.

that the expression of the BL6 allele of *Sox2* was completely restored, and expression of hmPIGA and BSD was undetectable (Fig. 4*G*).

To demonstrate the flexibility of Big-IN for delivery of payloads to additional loci, LP-PIGA2 was integrated into chromosome 7 of BL6xCAST *ΔPiga* mESCs, replacing a 157-kb region of the *Igf2/H19* locus (*SI Appendix*, Fig. S5*A*). We transfected these cells with pCAG-iCre and either the nonscarless payload pSox2^{46kb}-MC-BBTK or the scarless pSox2^{46kb} payload. Following stable positive selection with blasticidin and negative selection with proaerolysin and GCV, 95 of 96 (99%) of Sox2^{46kb}-MC clones were verified by PCR for the loss of LP-PIGA2 and the gain of the novel left payload junction (*SI Appendix*, Fig. S5*B*). Conversely, following transient blasticidin and proaerolysin selection, 12 of 48 (25%) Sox2^{46kb} clones were similarly verified. Further verification of selected clones confirmed the presence of the right payload junction for 24 of 25 clones and the absence of pCAG-iCre in all clones. Capture-seq analysis of chosen clones confirmed specific payload gain without detectable payload backbone and complete loss of LP-PIGA2 (*SI Appendix*, Fig. S5 *C* and *D*). Notably, Capture-seq analysis also identified clones with defects not easily detectable through PCR genotyping, including an internal payload duplication in BL6xCAST Sox2^{46kb}-MC clone C9 and an internal payload deletion in BL6xCAST Sox2^{46kb}-MC clone A4 (*SI Appendix*, Fig. S6).

Genomic Screening of On- and Off-Target Integrations. To screen genomic data for on- and off-target integration events, we developed *bamintersect*, which leverages a modular mapping approach where sequencing reads are mapped separately to two reference genomes. *Bamintersect* then jointly analyzes both mappings to detect read pairs indicative of a junction (Fig. 5*A*). Nearby reads in each reference are clustered and masked for uninformative regions (*Materials and Methods*). We applied *bamintersect* to confirm LP integration and payload delivery for the genomic engineering events described herein, the majority of which were verified by identifying multiple reads supporting the novel junctions between the integrated sequence and its flanks (Fig. 5 and Dataset S2).

For LP-PIGA integration at *Sox2* in BL6xCAST mESCs, two of the four analyzed clones (A1 and C5) were validated for the presence of both correct junctions, whereas one clone (C2) was validated only for the left junction, and an additional clone (G2) demonstrated off-target LP integration at chromosome 1 (Fig. 5*C*). *Bamintersect* also detected an unexpected junction between the right and left HAs for clones A1 and C5 (Dataset S2). PCR confirmed a tandem head-to-tail multimeric LP integration (*SI Appendix*, Fig. S7 *A* and *B*). All payloads delivered to clone A1 were verified as correctly targeted (Fig. 5 *B* and *D–H*) and lacked tandem LP junctions (*SI Appendix*, Fig. S7*C* and Dataset S2), suggesting the tandem LP supported productive recombination upon Cre expression.

Several junctions were impossible to confirm using *bamintersect* for technical reasons. For LP-TK integration at *HPRT1*, the 1-kb HAs precluded mapping reads spanning the junction between LP-TK and hg38. For PL1 deliveries to both *HPRT1* and *Sox2*, the left junction is nearly identical to that of the replaced LP. Although Sox2^{143kb} delivery results in junctions nearly identical to the CAST allele, the high rate of endogenous variation in BL6xCAST mESCs permitted specific detection of the correctly integrated payload. Analysis of read pairs overlapping informative BL6xCAST variants revealed that, while LP-PIGA mESCs junctions are depleted of BL6 reads, both junctions including the BL6 allele are restored in Sox2^{143kb} clone G11 mESCs (Fig. 5*F*). These results support the utility of *bamintersect* as a sensitive, scalable, and unbiased tool for detection of on and off-target integration events.

Discussion

We have described Big-IN, a platform for scalable targeted integration into mammalian genomes, and demonstrated its flexibility, efficiency, and precision at three loci in mouse and human

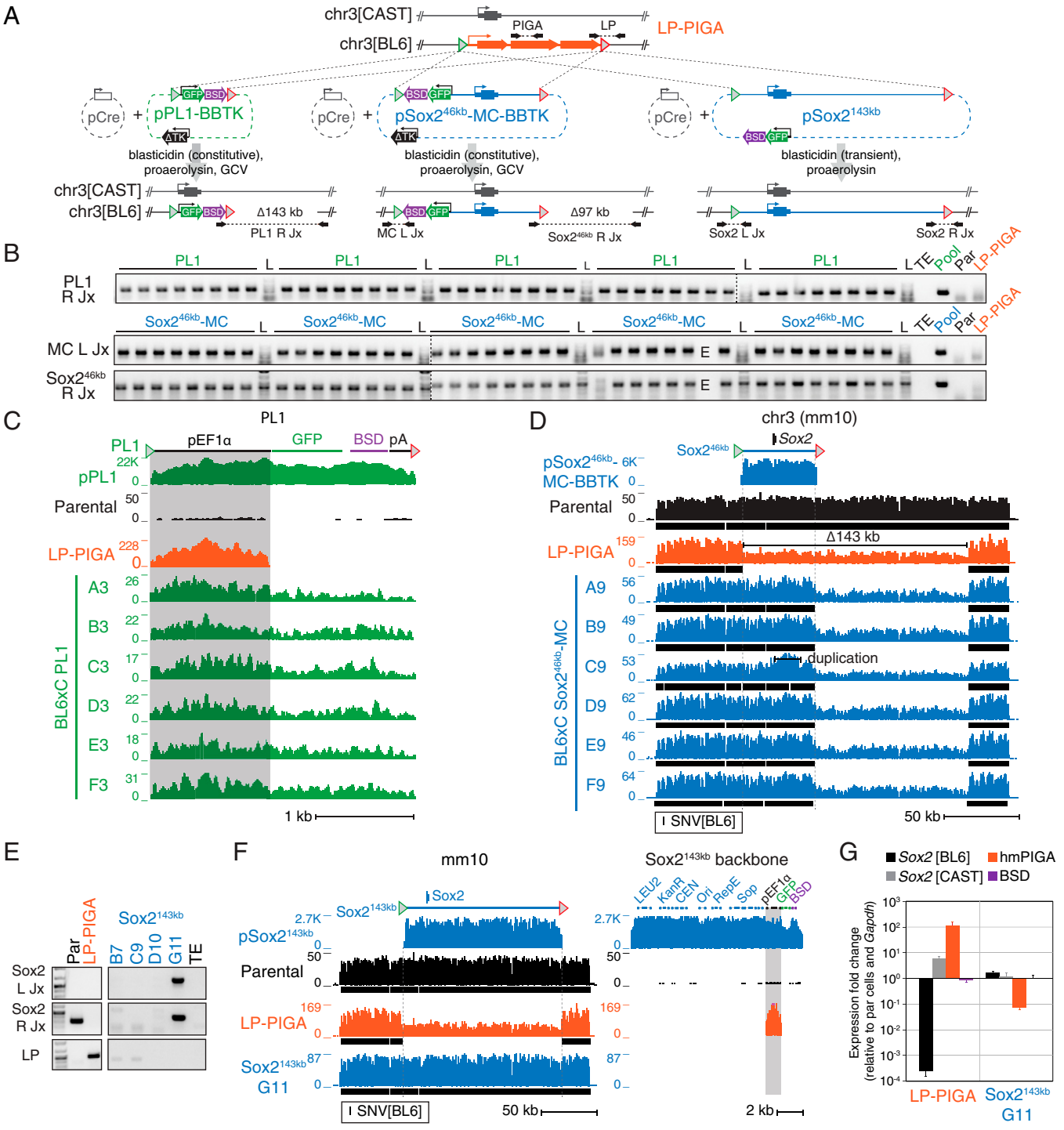

Fig. 4. Efficient delivery to mESCs. (*A*) Delivery of three payloads to BL6xCAST Δ*Piga* LP-PIGA mESCs. (*B*) PCR genotyping of PL1 (*Upper*) and Sox2^{46kb}-MC (*Lower*) mESC clones for novel junctions illustrated in *A*. E, empty well; L, ladder. (*C* and *D*) Capture-seq analysis of chosen PL1 and Sox2^{46kb}-MC mESC clones, with Parental and LP-PIGA mESCs as controls. (*C*) Sequencing coverage mapped to PL1. pEF1α (shaded gray) is present in both LP-PIGA and PL1. (*D*) Gain of coverage in Sox2^{46kb}-MC mESCs at the 46-kb payload region. Black ticks under each coverage track indicate detection of BL6 alleles at known SNVs. Internal payload duplication marked in Clone C9 (*SI Appendix*, Fig. S6). (*E*) PCR genotyping of Sox2^{143kb} clones for BL6-specific junctions and loss of LP-PIGA, as illustrated in *A*. (*F*) Sox2^{143kb} mESCs show restored coverage of the full 143-kb genomic region corresponding to the payload. Black ticks under each coverage track indicate detection of BL6 alleles at known SNVs. Coverage at right shows no retention of payload backbone. Cross-mapping sequences are shaded gray. (*G*) qRT-PCR expression analysis of Sox2^{143kb} clone G11 and LP-PIGA mESCs for mRNAs from BL6 and CAST *Sox2* alleles, payload-derived BSD, and LP-harbored hmPIGA. Bars represent mean + SD for technical replicates (n = 3).

ESCs. Big-IN first targets an LP to a locus of interest using CRISPR/Cas9-mediated HDR, which permits single-step payload integration through Cre-mediated RMCE (Fig. 6). Single-step payload integration minimizes confounding technical factors by permitting repeated deliveries to the same allele, and is thus ideal for in-depth interrogation of a given locus (47). LP cell lines can be

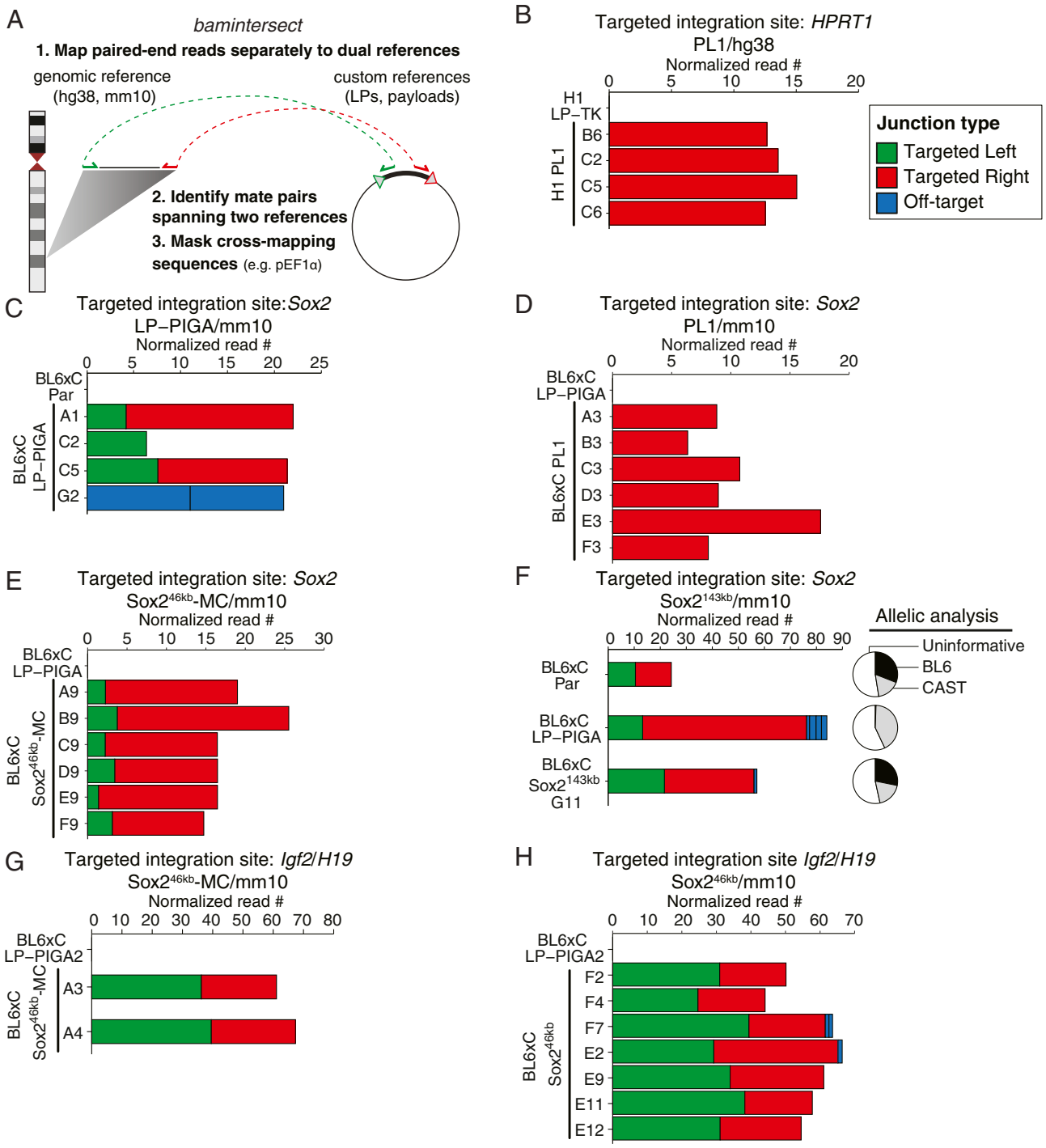

Fig. 5. *Bamintersect*, a tool for integration site analysis. (*A*) Schematic of the *bamintersect* analysis pipeline. (*B–H*) *Bamintersect* results between genomic and custom references indicated at top of each panel. Bars represent the number of reads supporting each junction, normalized to 10 million sequenced reads. Results were annotated as targeted left junction (green), targeted right junction (red), or off-target (blue). For PL1 integration at both *HPRT1* and *Sox2* (*B* and *D*), the targeted left junction is not distinguishable due to its near identity with the LP sequence being replaced. For integrations at *Sox2* (*C*, *E*, and *F*), the targeted left junction is adjacent to a low mappability region composed of simple repeats and an Alu sequence, consistently yielding fewer reads relative to the right junction. Allelic analysis in *F* categorizes reads at expected left and right junctions using known BL6xCAST SNVs; uninformative reads do not overlap known variants.

intensively verified following CRISPR/Cas9 expression to ensure the absence of undesired rearrangements or other off-target events, while subsequent Cre expression for payload delivery is expected to be less mutagenic (12).

Our cell-engineering approach is designed to scale rapidly across multiple loci and cell lines. While we have demonstrated Big-IN in both mouse and human ESCs, it is possible that engineering other mammalian cell lines with LPs may require optimization. Indeed, we note that despite the success of the LP-expressed Cre^{ERT2} strategy in H1 hESCs, exogenous Cre was required in mESCs. We have shown that the selection and delivery methods described herein can be redeployed in a modular

fashion to overcome challenges associated with different cell types and loci. For example, the LP can employ either HSV1-ΔTK or hmPIGA as a counterselectable marker, with the former suffering from a bystander effect, and the latter requiring prior engineering to inactivate the endogenous *PIGA*/*Piga* gene. While loss of GPI-anchored proteins has no detectable phenotype in culture, mice completely lacking *Piga* function are inviable (48). A reversible *Piga* knockout using an excisable intronic transcription terminator as previously engineered for *HPRT1* (49) would enable efficient recovery of *Piga*-expressing cells by sorting for a GPI-anchored membrane protein. A similar trade-off relates to the inclusion of a positive selection marker on the payload, which augments delivery efficiency, while its placement in the payload backbone enables scarless integration (*SI Appendix*, Fig. S5*B*). Quantitative comparison of the efficiency of the Big-IN deliveries described herein is confounded by technical differences and the need to replate rapidly growing ESCs, but we expect that future improvements will enhance overall efficiency and its application to diverse cellular contexts.

Our verification strategy is tailored to enable early verification of engineering outcomes. For example, the use of locally generated Capture-seq bait circumvents the cost and delay of commercially synthesized bait pools. Additionally, *bamintersect* works with standard libraries generated from genomic DNA, unlike specialized ligation-mediated approaches, and uses standard reference coordinates rather than custom assemblies for each delivery. We demonstrate the value of our pipeline through detection of tandem LP insertions and internal duplications and deletions in integrated payloads that would have been difficult to detect using PCR screening.

The effectiveness of Big-IN for integration of large DNA constructs (*SI Appendix*, Fig. S5*B*) suggests that it might also be optimized to support integration of complex libraries for saturation mutagenesis of shorter elements (50–52), and eventually, analysis of large constructs in a pooled library format. When combined with the rapidly evolving big DNA synthesis field (14, 25), we envision that Big-IN will enable designer-like control over mammalian genomes and facilitate a synthetic approach to genome biology.

Materials and Methods

Additional information is available in *SI Appendix*.

gRNA Design. gRNAs were designed using the GuideScan algorithm (53). For allele-specific LP integration at *Sox2*, we produced a scored list of potential gRNAs targeting a 261-kb region surrounding *Sox2* using the BL6 reference genome sequence. Next, we identified gRNAs for which the corresponding PAM is mutated in the CAST allele, resulting in a list of BL6-specific gRNAs. From this list we selected two high-scoring gRNAs, Sox2-g1 and Sox2-g2, which target a 143-kb genomic region for replacement with the LP. gRNA sequences are listed in *SI Appendix*, Table S2.

Cell Culture. WA01 (H1) hESCs were purchased from WiCell. The use of H1 hESCs was approved by the New York University School of Medicine Embryonic Stem Cell Research Oversight Committee. H1 hESCs were initially grown for 2 wk on plates coated with Matrigel (Corning 354277) in mTeSR medium (Stem Cell Technologies 85850) and subsequently transferred to plates coated with Geltrex (Gibco A1413302) and StemFlex medium (ThermoFisher A3349401) supplemented with 1% Pen-Strep (ThermoFisher 15140122). For routine passaging, cells were dissociated into clumps with Versene (Gibco 15-040-066) and gentle trituration. Wide-orifice pipette tips were used when handling small volumes of cell suspension.

C57BL6/6J × CAST/EiJ (BL6xCAST) clone 4 mESCs (41) were kindly provided by David Spector, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY. mESCs were cultured on plates coated with 0.1% gelatin (EMD Millipore ES-006-B) in 80/20 medium comprising 80% 2i medium and 20% mESC medium. 2i medium contained a 1:1 mixture of Advanced DMEM/F12 (ThermoFisher 12634010) and Neurobasal-A (ThermoFisher 10888022) supplemented with 1% N2 Supplement (ThermoFisher 17502048), 2% B27 Supplement (ThermoFisher 17504044), 1% glutamax (ThermoFisher 35050061), 1% Pen-Strep (ThermoFisher 15140122), 0.1 mM 2-mercaptoethanol (Sigma M3148), 1,250 U/mL LIF (ESGRO ESG1107l), 3 μM CHIR99021 (R&D Systems 4423), and 1 μM PD0325901 (Sigma PZ0162). mESC medium contained knockout DMEM (ThermoFisher 10829018) supplemented with 15% FBS (BenchMark 100-106), 0.1 mM 2-mercaptoethanol, 1% glutamax, 1% MEM nonessential amino acids (ThermoFisher 11140050), 1% nucleosides (EMD Millipore ES-008-D), 1% Pen-Strep, and 1,250 U/mL LIF. HEK-293T cells were cultured in DMEM supplemented with 10% FBS, 1 mM sodium pyruvate (ThermoFisher 11360070), 1% glutamax, and 1% Pen-Strep. All cells were grown at 37 °C in a humidified atmosphere of 5% CO_2 and passaged on average twice per week.

Genome Engineering. Relevant genomic coordinates are listed in *SI Appendix*, Table S3.

H1 hESCs were transfected using the Neon Transfection System (ThermoFisher). Cells were treated several hours prior to transfection with StemFlex medium supplemented with 1% RevitaCell Supplement (ThermoFisher A2644501). Cells were washed with PBS, dissociated into a single-cell suspension using TrypLE-Select (ThermoFisher 12563011), which was neutralized with StemFlex medium, spun down at 200 relative centrifugal force

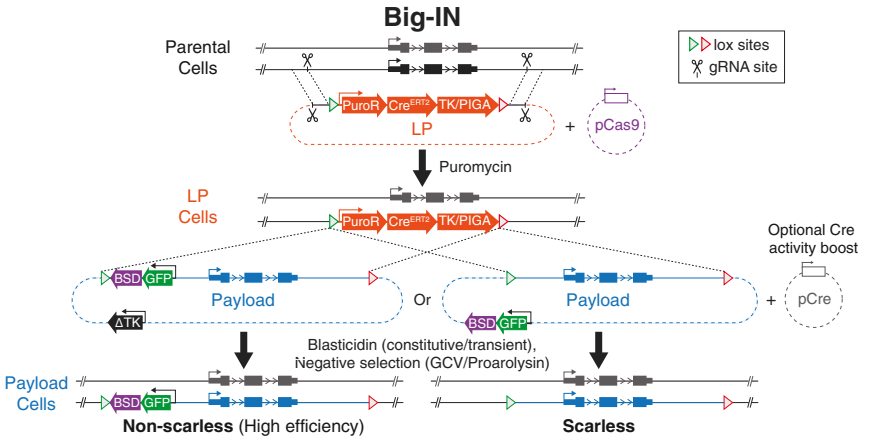

Fig. 6. Targeted locus-scale genome rewriting using Big-IN. An allele of interest is replaced by a LP using CRISPR/Cas9-mediated HDR. A pair of gRNAs target the termini of the replaced allele and the LP, and short HAs mediate precise LP integration. Puromycin selects for LP-harboring cells. Next, Cre-mediated recombination of two pairs of heterotypic loxM and loxP sites results in LP/payload cassette exchange and resistance to either GCV for HSV1-ΔTK LPs, or proaerolysin for hmPIGA LPs in cells where endogenous PIGA is inactivated. Positioning the blasticidin cassette (BSD) within the payload permits election for high-efficiency integration; positioning BSD on the payload backbone permits transient selection for scarless delivery. Additionally, backbone HSV-ΔTK (*Left*) can be counterselected with GCV to limit off-target integration. Each engineering step is comprehensively verified by PCR genotyping, WGS or Capture-seq, and functional assays.

(rcf) for 3 min, supernatant aspirated, and cells resuspended in PBS. Next, 1×10^6 cells per transfection were spun down at 200 rcf for 3 min and resuspended in Neon Buffer R at a final concentration of 2×10^7 cells/mL; 50 μL of cell suspension were mixed with 50 μL Neon Buffer R containing 10 μg of total DNA per transfection. Nucleofection used Neon 100 μL Tips with two 20-ms pulses at 1,100 V. Transfected cells were transferred into plates coated with rhLaminin-521 (Gibco A29249) prefilled with StemFlex medium supplemented with 1% RevitaCell. *PIGA* deletion was performed with 5 μg of each pCas9 plasmid expressing gRNAs hPIGA-g1 and hPIGA-g2 and cells were selected with 200 pM proaerolysin for 1 to 2 wk posttransfection. These Δ*PIGA* cells were used for subsequent LP-PIGA integrations. All LP integrations at *HPRT1* were performed using 5 μg of the pLP and 2.5 μg of each pCas9 plasmid expressing HPRT1-g1 and HPRT1-g2 gRNAs, and cells were selected using a combination of 1 μg/mL puromycin and 6-TG, as indicated. H1 PL1 integrations were performed using 5 μg pPL1. Cells were treated with 200 nM 4-hydroxytamoxifen (Tam) the day following transfection for 3 h, selected with 5 μg/mL blasticidin S for 8 d, followed by 4 d of selection with 100 nM GCV to eliminate TK-expressing cells.

LP integrations and genomic deletions in BL6xCAST mESCs were performed using the Neon Transfection System. Cells were washed with PBS, dissociated into a single-cell suspension using TrypLE-Select (Gibco), which was neutralized with mESC medium, spun down at 200 rcf for 3 min, supernatant aspirated, and cells resuspended in PBS. Next 1×10^6 cells per transfection were spun down at 200 rcf for 3 min and resuspended in Neon Buffer R at a final concentration of 2×10^7 cells/mL. Per transfection, 50 μL of cell suspension were mixed with 50 μL Neon Buffer R containing 10 μg of total DNA and nucleofected using Neon 100 μL tips with two 20-ms pulses at 1200 V. Transfected cells were transferred into gelatin-coated plates prefilled with 80/20 medium. *Piga* deletion was performed with 5 μg of each pCas9 plasmid expressing gRNAs mPiga-g1 and mPiga-g2, and cells were selected with 2 nM proaerolysin ~1 wk posttransfection. Δ*Piga* cells were used for subsequent LP integrations. LP-PIGA integrations at *Sox2* were performed using 5 μg of the pLP and 2.5 μg of each pCas9 plasmid expressing Sox2-g1 and Sox2-g2 gRNAs, and cells were selected with 1 μg/mL puromycin. LP-PIGA2 integration at *Igf2/H19* was performed using 5 μg of the pLP-PIGA2 and 2.5 μg of each pCas9 plasmid expressing Igf2/H19-g1 and Igf2/H19-g2 gRNAs, and cells were selected with 1 μg/mL puromycin followed by selection with 1 μM GCV.

Payload deliveries in BL6xCAST mESCs were performed using a Nucleofector 2b (Lonza). Cells were washed with PBS, dissociated into a single-cell suspension using TrypLE-Select, which was neutralized with mESC medium, spun down at 200 rcf for 3 min, supernatant aspirated, and cells resuspended in ice-cold PBS, counted, and 5×10^6 cells per transfection were spun down at 200 rcf for 3 min and resuspended in a room temperature mixture of 82 μL nucleofector solution and 18 μL nucleofector supplement from the Mouse ES Cell Nucleofector kit (Lonza VPH-1001). Per transfection, 100 μL of cell suspension were mixed with 10 μL TE containing 2.25 to 5 μg of total DNA, and nucleofected using program A-23. PL1 deliveries were performed with 1.5 μg pPL1-BBTK and 0.75 μg pCAG-Cre (Addgene plasmid #13775). pSox2^{46kb}-MC deliveries (failed deliveries) were performed with 35 μg pSox2^{46kb}-MC. Payload-transfected mESCs were treated with 200 nM Tam for 4 h before and 24 h after transfection. Cells were selected with blasticidin constitutively starting day 1 posttransfection and with 2 nM proaerolysin for 2 d starting day 14 posttransfection. pSox2^{46kb}-MC-BBTK deliveries were performed with 3 μg pSox2^{46kb}-MC-BBTK and 1 μg pCAG-Cre. Payload-transfected mESCs were treated with 200 nM Tam for 24 h before and after transfection. mESCs were grown for 10 d with blasticidin. On days 11 and 12, 1 nM proaerolysin was added, and on days 13 and 14, 1 μM GCV was also added. pSox2^{143kb} delivery was performed with 0.3 μg pSox2^{143kb} and 2 μg pCAG-iCre (Addgene plasmid #89573). Payload-transfected mESCs were selected with blasticidin for 2 d starting day 1 posttransfection and with 2 nM proaerolysin for 2 d starting day 7 posttransfection. Payload deliveries to BL6xCAST *Igf2/H19* were performed with 5 μg pSox2^{46kb}-MC-BBTK or pSox2^{46kb} and 2 μg pCAG-iCre. Cells were selected with blasticidin either transiently during days 1 and 2 posttransfection (pSox2^{46kb}) or constitutively (pSox2^{46kb}-MC-BBTK), followed by 2 nM proaerolysin selection during days 7 and 8 posttransfection. pSox2^{46kb}-MC-BBTK transfected cells were further selected with 1 μM GCV during days 9 and 10 posttransfection.

Preparation of Illumina Double-Stranded DNA Libraries. Genomic DNA was isolated from cells using the DNeasy Blood and Tissue Kit (Qiagen) according to the manufacturer's protocol. For this, 1,000 ng of DNA was sheared to ~500 to 900 bp in a 96-well microplate using the Covaris LE220 (450 W, 10% Duty Factor, 200 cycles per burst, and 90-s treatment time). Sheared DNA was purified using the DNA Clean and Concentrate-5 Kit (Zymo Research), and the concentration was measured on a Nanodrop instrument (Invitrogen). DNA fragments were end-repaired with T4 DNA polymerase, Klenow DNA polymerase, and T4 polynucleotide kinase (New England Biolabs), and A-tailed using Klenow (3′-5′ exo-; New England Biolabs). Illumina-compatible adapters were subsequently ligated to DNA ends, and DNA libraries were amplified with KAPA 2× Hi-Fi Hotstart Readymix (Roche). Sequenced samples are listed in Dataset S1.

Targeted Resequencing Using Capture-Seq. Baits for sequence capture were prepared from BAC or plasmid DNA containing the sequence of interest. BAC coordinates are listed in *SI Appendix*, Table S3. Biotin-16-dUTP (Roche) was incorporated into bait DNA using a Nick Translation kit (Roche). The reaction (total volume 20 μL) was set-up in a 200-μL PCR tube on ice as follows: 2 μg of BAC DNA, 10 μL of 0.1 mM Biotin-dUTP/dNTP mixture (1 volume Biotin-16-dUTP, 2 volumes dTTP, 3 volumes dATP, 3 volumes dCTP, and 3 volumes dGTP), 2 μL of 10× nick translation buffer, and 2 μL of enzyme mixture. Nick translation was carried out at 15 °C for 16 h or 8 h (for BAC or plasmid DNA, respectively) in a thermal cycler. The reaction was stopped by addition of 1 μL 0.5 M EDTA and heating at 65 °C for 10 min or cooling at 4 °C overnight. Biotinylated baits were purified by ethanol precipitation, resuspended in 50 mL H_2O, and the concentration was measured on a Nanodrop instrument. Baits were stored at −20 °C.

Targeted sequencing using in-solution hybridization capture (Capture-seq) was performed as described previously (54), with modifications. One microgram biotinylated DNA bait and 10 μg Cot-1 human or mouse DNA (Invitrogen) were combined with universal and sample-specific blocking oligos and lyophilized using a SpeedVac. Lyophilized DNA was resuspended in 12 μL TE (pH 7.5) and overlaid with mineral oil. In a thermal cycler, the DNA mixture was denatured at 96 °C for 5 min, incubated at 65 °C for an additional 15 min, and then 12 μL of 2× hybridization buffer (1.5 M NaCl, 40 mM sodium phosphate buffer [pH 7.2], 10 mM EDTA [pH 8], 10× Denhardt's, and 0.2% SDS) was added to the DNA, and the mixture was prehybridized for 6 h at 65 °C.

A total of 1 μg from up to two to eight libraries were pooled into a single 200-μL PCR tube for a single-capture reaction. Library DNA was diluted in H_2O to a final volume of 12 μL and overlaid with mineral oil. Library DNA was denatured at 96 °C for 5 min, incubated at 65 °C for an additional 15 min, and then 12 μL of 2× hybridization buffer was added to the denatured DNA library. The entire volume (24 μL) of denatured library DNA was added to the tube of prehybridized bait DNA, and the mixture was incubated at 65 °C for 16 to 22 h. For each capture reaction, 50 μL of MyOne streptavidin-coated magnetic beads (Invitrogen) were washed with 1× B&W buffer (5 mM Tris·HCl pH 7.5, 0.5 mM EDTA, 1 M NaCl) three times, and then resuspended in 150 μL 1× B&W buffer in a low-retention microcentrifuge tube. The hybridization mix (48 μL) plus 48 μL 2× B&W buffer (10 mM Tris·HCl pH 7.5, 1 mM EDTA, 2 M NaCl) were then combined with the prewashed magnetic beads, and incubated at room temperature for 30 min with rotation. The magnetic beads were washed once at 25 °C for 15 min in 1× SSC with 0.1% SDS and three times at 65 °C for 15 min in 0.1× SSC with 0.1% SDS. To denature the captured library DNA, the beads were resuspended in 100 μL 100 mM NaOH, and incubated at room temperature for 10 min. After allowing the beads to separate on a magnetic rack, the supernatant (containing enriched library DNA) was transferred to a new tube, neutralized with 100 μL 1 M Tris·HCl pH 7.5, and purified using the DNA Clean and Concentrate-5 Kit (Zymo Research). Four microliters of the captured library DNA were evaluated using qPCR to determine the optimal number of final PCR amplification cycles. Captured libraries were then amplified with KAPA Hi-Fi Hotstart Readymix (Roche). Bait sets and sequencing statistics are listed in Dataset S1.

Sequencing Data Processing. Illumina libraries were sequenced in paired-end mode on an Illumina NextSeq. 500 operated at the Institute for Systems Genetics or a NovaSeq. 6000 operated by the New York University Langone Health Genome Technology Center. Reads were demultiplexed with Illumina bcl2fastq v2.20 requiring a perfect match to indexing BC sequences. All WGS and Capture-seq data were processed using a uniform mapping and peak calling pipeline. Illumina sequencing adapters were trimmed with Trimmomatic v0.39 (55). Sequencing reads were aligned using BWA v0.7.17 (56) to a reference genome (GRCh38/hg38 or GRCm38/mm10), including unscaffolded contigs and alternate references, as well as independently to custom references for relevant vectors. PCR duplicates were marked using samblaster v0.1.24 (57). Generation of per base coverage depth tracks and quantification was performed using BEDOPS v2.4.35 (58). Data were visualized using the University of California, Santa Cruz Genome Browser. The sequencing processing pipeline is available at https://github.com/mauranolab/mapping.

Genotype Analysis. Variant calling was performed on sequenced BL6xCAST samples to verify correct allele-specific engineering using a standard pipeline based on bcftools v1.9:

bcftools mpileup–redo-BAQ–adjust-MQ 50–gap-frac 0.05–max-depth 10000–max-idepth 200000 -a DP,AD–output-type u |

bcftools call–keep-alts –ploidy 1–multiallelic-caller -f GQ–output-type u

Raw pileups were filtered using:

bcftools norm–check-ref w–output-type u |

bcftools filter -i "INFO/DP>=10 & QUAL>=10 & GQ>=99 & FORMAT/DP>=10"–SnpGap 3–IndelGap 10–set-GTs .–output-type u |

bcftools view -i 'GT="alt"'–trim-alt-alleles–output-type z

SNVs called in each sample were intersected with expected BL6/CAST heterozygous sites based on known variants called for CAST/EiJ (42).

Analysis of Integration Junctions Using *Bamintersect*. *Bamintersect* enables efficient filtering paired-end genomic sequencing based on dual independent mapping to different references, typically a mammalian reference genome (hg38 or mm10) and an engineered reference of interest (LP or payload). *Bamintersect* identifies junctions through analysis of read pairs where each read is mapped to a different reference. For LP/payload genomes, the read's mate is required to be unmapped to that genome. Reads must be fully mapped with ≤1 mismatched base and no clipping, insertions, or deletions, and duplicate or supplementary alignments are excluded. *Bamintersect* additionally assesses informative junctions from two further classes of read pairs mapping to the same reference: 1) Between the LP/payload and vector backbone; and 2) between the HAs and other genomic regions.

Masking is applied to each bam file to eliminate read pairs mapping to sequences present in multiple contexts (>120 bps with >85% identity): 1) hmPIGA, human *EIF1* poly(A), ERT2, and pEF1α (LPs); 2) the human U6 promoter (pCas9); and 3) pEF1α (payload deliveries). Satellite repeats were similarly masked. Reported reads spanning references were required to include a minimum of 20-bp mapping outside additional filtered regions: 1) HAs (LP integrations), 2) lox sites and genomic sequence corresponding to the payload (payload deliveries), and 3) the intervening deleted region (HA analyses for payload deliveries).

For all analyses, reads with the same strand and mapping to within 500 bp of each other were clustered for reporting. Regions below 75 bp or with fewer than 1 read/10M reads sequenced were excluded. A distance 1 kbp or greater was required between regions mapping to the same chromosome.

Data Availability. The sequencing processing pipeline and genome browser visualization hub source code are available at https://github.com/mauranolab/mapping. Sequencing data have been deposited in the Gene Expression Omnibus (GEO) database, https://www.ncbi.nlm.nih.gov/geo (accession no. GSE159488).

ACKNOWLEDGMENTS. We thank Dr. David J. Araten, Dr. Dongui Li, Dr. Paolo Mita, Brendan Camellato, and Julie Trolle for providing reagents and advice. This work was partially funded by NIH Grants RM1HG009491 (to J.D.B.) and R35GM119703 (to M.T.M.), and by the Colton Center for Autoimmunity.

1. M. T. Maurano *et al*., Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
2. R. D. Palmiter, R. L. Brinster, Germ-line transformation of mice. *Annu. Rev. Genet.* **20**, 465–499 (1986).
3. A. Battle, C. D. Brown, B. E. Engelhardt, S. B. Montgomery; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; eQTL manuscript working group, Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
4. M. T. Maurano *et al*., Identification of cellular context sensitive regulatory variation in mouse genomes. bioRxiv:2020.06.27.175422 (28 June 2020).
5. O. Smithies, R. G. Gregg, S. S. Boggs, M. A. Koralewski, R. S. Kucherlapati, Insertion of DNA sequences into the human chromosomal beta-globin locus by homologous recombination. *Nature* **317**, 230–234 (1985).
6. K. R. Thomas, K. R. Folger, M. R. Capecchi, High frequency targeting of genes to specific sites in the mammalian genome. *Cell* **44**, 419–428 (1986).
7. F. D. Urnov, B. C. Genome Editing, Genome editing B.C. (Before CRISPR): Lasting lessons from the "Old Testament". *CRISPR J.* **1**, 34–46 (2018).
8. J. Vierstra *et al*., Functional footprinting of regulatory DNA. *Nat. Methods* **12**, 927–930 (2015).
9. N. E. Sanjana *et al*., High-resolution interrogation of functional elements in the noncoding genome. *Science* **353**, 1545–1549 (2016).
10. M. Osterwalder *et al*., Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* **554**, 239–243 (2018).
11. A. Despang *et al*., Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271 (2019).
12. K. Boroviak, B. Fu, F. Yang, B. Doe, A. Bradley, Revealing hidden complexities of genomic rearrangements generated with Cas9. *Sci. Rep.* **7**, 12867 (2017).
13. S. M. Richardson *et al*., Design of a synthetic yeast genome. *Science* **355**, 1040–1044 (2017).
14. W. Zhang, L. A. Mitchell, J. S. Bader, J. D. Boeke, Synthetic genomes. *Annu. Rev. Biochem.* **89**, 77–101 (2020).
15. N. Heintz, BAC to the future: The use of bac transgenic mice for neuroscience research. *Nat. Rev. Neurosci.* **2**, 861–870 (2001).
16. K. R. Peterson *et al*., Transgenic mice containing a 248-kb yeast artificial chromosome carrying the human beta-globin locus display proper developmental control of human globin genes. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 7593–7597 (1993).
17. A. Schedl, L. Montoliu, G. Kelsey, G. Schütz, A yeast artificial chromosome covering the tyrosinase gene confers copy number-dependent expression in transgenic mice. *Nature* **362**, 258–261 (1993).
18. K. R. Peterson *et al*., Use of yeast artificial chromosomes (YACs) in studies of mammalian development: Production of beta-globin locus YAC mice carrying human globin developmental mutants. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 5655–5659 (1995).
19. J. Seibler, D. Schübeler, S. Fiering, M. Groudine, J. Bode, DNA cassette exchange in ES cells mediated by Flp recombinase: An efficient strategy for repeated modification of tagged loci by marker-free constructs. *Biochemistry* **37**, 6229–6234 (1998).
20. E. E. Bouhassira, K. Westerman, P. Leboulch, Transcriptional behavior of LCR enhancer elements integrated at the same chromosomal locus by recombinase-mediated cassette exchange. *Blood* **90**, 3332–3344 (1997).
21. M. Iacovino *et al*., Inducible cassette exchange: A rapid and efficient system enabling conditional gene expression in embryonic stem and primary cells. *Stem Cells* **29**, 1580–1588 (2011).
22. F. Zhu *et al*., DICE, an efficient system for iterative genomic editing in human pluripotent stem cells. *Nucleic Acids Res.* **42**, e34 (2014).
23. K. A. Matreyek, J. J. Stephany, D. M. Fowler, A platform for functional assessment of large variant libraries in mammalian cells. *Nucleic Acids Res.* **45**, e102 (2017).
24. H. A. Wallace *et al*., Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. *Cell* **128**, 197–209 (2007).
25. L. A. Mitchell *et al*., De novo assembly, delivery and expression of a 101 kb human gene in mouse cells. bioRxiv:10.1101/423426 (26 January 2019).
26. L. A. Pennacchio *et al*., In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
27. M. E. Dickinson *et al*.; International Mouse Phenotyping Consortium; Jackson Laboratory; Infrastructure Nationale PHENOMIN, Institut Clinique de la Souris (ICS); Charles River Laboratories; MRC Harwell; Toronto Centre for Phenogenomics; Wellcome Trust Sanger Institute; RIKEN BioResource Center, High-throughput discovery of novel developmental phenotypes. *Nature* **537**, 508–514 (2016).
28. M. H. St Clair, C. U. Lambe, P. A. Furman, Inhibition by ganciclovir of cell growth and DNA synthesis of cells biochemically transformed with herpesvirus genetic information. *Antimicrob. Agents Chemother.* **31**, 844–849 (1987).
29. R. H. Friedel, W. Wurst, B. Wefers, R. Kühn, Generating conditional knockout mice. *Methods Mol. Biol.* **693**, 205–231 (2011).
30. M. D. Ryan, A. M. King, G. P. Thomas, Cleavage of foot-and-mouth disease virus polyprotein is mediated by residues located within a 19 amino acid sequence. *J. Gen. Virol.* **72**, 2727–2732 (1991).
31. C. T. Caskey, G. D. Kruh, The HPRT locus. *Cell* **16**, 1–9 (1979).
32. F. A. Ran *et al*., Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
33. X. Yao *et al*., Tild-CRISPR allows for efficient and precise gene knockin in mouse and human cells. *Dev. Cell* **45**, 526–536.e5 (2018).
34. C. Fillat, M. Carrió, A. Cascante, B. Sangro, Suicide gene therapy mediated by the herpes simplex virus thymidine kinase gene/ganciclovir system: Fifteen years of application. *Curr. Gene Ther.* **3**, 13–26 (2003).
35. A. A. Elshami *et al*., Gap junctions play a role in the 'bystander effect' of the herpes simplex virus thymidine kinase/ganciclovir system in vitro. *Gene Ther.* **3**, 85–92 (1996).
36. M. Mesnil, C. Piccoli, G. Tiraby, K. Willecke, H. Yamasaki, Bystander killing of cancer cells by herpes simplex virus thymidine kinase gene is mediated by connexins. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 1831–1835 (1996).
37. Y. Iida *et al*., Characterization of genomic PIG-A gene: A gene for glycosylphosphatidylinositol-anchor biosynthesis and paroxysmal nocturnal hemoglobinuria. *Blood* **83**, 3126–3131 (1994).
38. D. B. Diep, K. L. Nelson, S. M. Raja, E. N. Pleshak, J. T. Buckley, Glycosylphosphatidylinositol anchors of membrane glycoproteins are binding determinants for the channel-forming toxin aerolysin. *J. Biol. Chem.* **273**, 2355–2360 (1998).

39. D. J. Araten, K. Nafa, K. Pakdeesuwan, L. Luzzatto, Clonal populations of hematopoietic cells with paroxysmal nocturnal hemoglobinuria genotype and phenotype are present in normal individuals. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 5209–5214 (1999).
40. D. Li *et al*., Application of counter-selectable marker PIGA in engineering designer deletion cell lines and characterization of CRISPR deletion efficiency. *Nucleic Acids Res.*, 10.1093/nar/gkab035 (2021).
41. M. A. Eckersley-Maslin *et al*., Random monoallelic gene expression increases upon embryonic stem cell differentiation. *Dev. Cell* **28**, 351–365 (2014).
42. T. M. Keane *et al*., Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
43. A. A. Avilion *et al*., Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev.* **17**, 126–140 (2003).
44. H. Y. Zhou *et al*., A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev.* **28**, 2699–2711 (2014).
45. Y. Li *et al*., CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One* **9**, e114485 (2014).
46. D. S. Bindels *et al*., mScarlet: A bright monomeric red fluorescent protein for cellular imaging. *Nat. Methods* **14**, 53–56 (2017).
47. J. M. Laurent *et al*., Big DNA as a tool to dissect an age-related macular degeneration-associated haplotype. *Precis. Clin. Med.* **2**, 1–7 (2019).
48. K. Kawagoe *et al*., Glycosylphosphatidylinositol-anchor-deficient mice: Implications for clonal dominance of mutant cells in paroxysmal nocturnal hemoglobinuria. *Blood* **87**, 3600–3606 (1996).
49. Y. Han *et al*., Orientation-dependent regulation of integrated HIV-1 expression by host gene transcriptional readthrough. *Cell Host Microbe* **4**, 134–146 (2008).
50. D. M. Fowler, S. Fields, Deep mutational scanning: A new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
51. E. Z. Kvon *et al*., Comprehensive in vivo interrogation reveals phenotypic impact of human enhancer variants. *Cell* **180**, 1262–1271.e15 (2020).
52. B. B. Maricque, H. G. Chaudhari, B. A. Cohen, A massively parallel reporter assay dissects the influence of chromatin structure on *cis*-regulatory activity. *Nat. Biotechnol.* **37**, 90–95 (2018).
53. A. R. Perez *et al*., GuideScan software for improved single and paired CRISPR guide RNA design. *Nat. Biotechnol.* **35**, 347–349 (2017).
54. E. Yigit *et al*., High-resolution nucleosome mapping of targeted regions using BAC-based enrichment. *Nucleic Acids Res.* **41**, e87 (2013).
55. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
56. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
57. G. G. Faust, I. M. Hall, SAMBLASTER: Fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505 (2014).
58. S. Neph *et al*., BEDOPS: High-performance genomic feature operations. *Bioinformatics* **28**, 1919–1920 (2012).
59. R. Feil, J. Wagner, D. Metzger, P. Chambon, Regulation of Cre recombinase activity by mutated estrogen receptor ligand-binding domains. *Biochem. Biophys. Res. Commun.* **237**, 752–757 (1997).