

# Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements

Michael Kosicki, Kärt Tomberg & Allan Bradley

**CRISPR–Cas9 is poised to become the gene editing tool of choice in clinical contexts. Thus far, exploration of Cas9-induced genetic alterations has been limited to the immediate vicinity of the target site and distal off-target sequences, leading to the conclusion that CRISPR–Cas9 was reasonably specific. Here we report significant on-target mutagenesis, such as large deletions and more complex genomic rearrangements at the targeted sites in mouse embryonic stem cells, mouse hematopoietic progenitors and a human differentiated cell line. Using long-read sequencing and long-range PCR genotyping, we show that DNA breaks introduced by single-guide RNA/Cas9 frequently resolved into deletions extending over many kilobases. Furthermore, lesions distal to the cut site and crossover events were identified. The observed genomic damage in mitotically active cells caused by CRISPR–Cas9 editing may have pathogenic consequences.**

The utility of the CRISPR–Cas9 system for gene therapy in humans has been recognized and extensively investigated<sup>1</sup>. Initial concerns about the off-target activity have been addressed by the development of sensitive detection methods, as well as modified Cas9 enzymes and improved delivery protocols that limit this type of damage<sup>2–12</sup>. The vast majority of on-target DNA repair outcomes after Cas9 cutting in a variety of cell types are thought to be insertions and deletions (indels) of less than 20 bp<sup>13–15</sup>. Although indels a few hundred nucleotides in size were also observed in experiments using Cas9 or other nucleases, they were reported to be rare<sup>16–18</sup>. Consequently, Cas9 has been assumed to be reasonably specific and the first approved clinical trials using Cas9 edited cells are underway (clinicaltrials.gov: [NCT03081715](#), [NCT03398967](#), [NCT03166878](#), [NCT02793856](#), [NCT03044743](#), [NCT03164135](#)).

Studies using paired gRNAs to induce localized deletions also reported generation of more complex genotypes, such as inversions, endogenous and exogenous DNA insertions, and larger-than-expected deletions<sup>19–23</sup>. Single gRNAs were shown to induce deletions of up to 600 bp in mouse zygotes<sup>24</sup>. Deletions of up to 1.5 kb in a haploid cancer cell line potentially induced by single gRNAs have been described, but since the guides were directed to a small part of the genome and provided as a pool, the possibility of

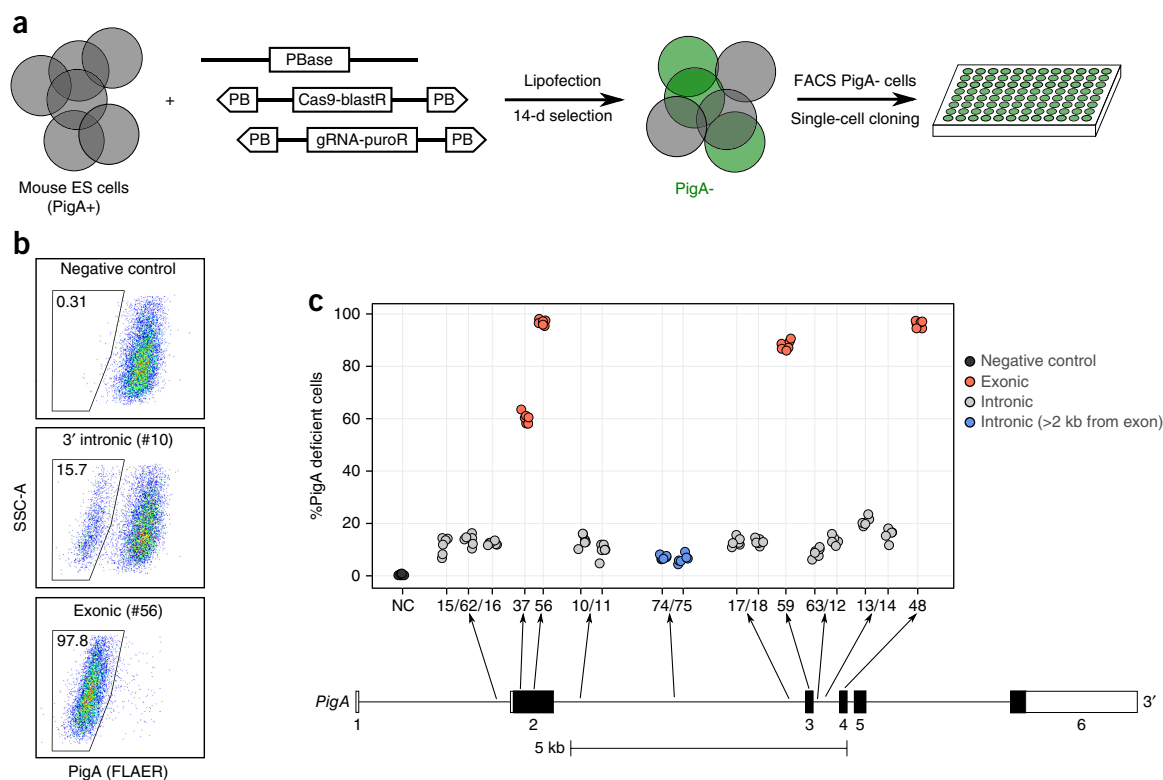
rare double-cutting events cannot be excluded<sup>25</sup>. Furthermore, the analysis of the alleles generated using both single and paired gRNAs has in most studies relied on amplification of short regions (<1 kb) around the target and potential off-target sites, limiting the scope of assessment. Lesions non-contiguous with the cleavage site, such as those reported in yeast upon *I-SceI* nuclease cutting, would also be missed by such short-range assessments<sup>26–28</sup>. Finally, cancer cell lines, whose genome and DNA repair mechanisms are abnormal, were often used in the context of studying Cas9-induced lesions, making extrapolations to normal tissues and cells problematic.

We speculate that current assessments may have missed a substantial proportion of potential genotypes generated by on-target Cas9 cutting and repair, some of which may have potential pathogenic consequences following somatic editing of large populations of mitotically active cells.

We first comprehensively explored allelic diversity induced by Cas9 at the X-linked *PigA* locus, which is hemizygous in male embryonic stem (ES) cells. In contrast to cancer-derived cell lines, ES cells have a normal karyotype and intact DNA repair mechanisms, which makes them more representative of a normal somatic cell. Although mouse ES cells and embryonic fibroblasts differ in their use of DNA repair pathways, it is not known how they compare to other somatic cells<sup>29</sup>. We introduced Cas9 and gRNA constructs targeting intronic and exonic sites of *PigA* into JM8 mouse ES cells using *PiggyBac* transposition. Cells with both constructs were selected and subsequently stained with FLAER reagent to quantify the proportion of *PigA*-deficient cells (**Fig. 1a,b**). Single gRNAs targeting exons 2 to 4 yielded very high rates of *PigA* loss (59–97%). Notably, single gRNAs targeting intronic sites also yielded *PigA*-deficient cells at significant frequencies. Ten different guides located 263–520 bp from the nearest exon caused 8–20% *PigA* loss, whereas two guides greater than 2 kb away induced 5–7% loss (**Fig. 1c** and **Supplementary Table 1**). We obtained similar results with transient expression using electroporation or lipofection of ribonucleoprotein complexes (RNP), proving that these observations were not a consequence of *PiggyBac* transposition, delivery method, antibiotic selection or cellular response to transfected plasmid DNA (**Supplementary Fig. 1**). Lower knockout efficiency using exonic guides correlated with slower editing dynamics when delivered by *PiggyBac* transposition (data not shown).

Wellcome Sanger Institute, Hinxton, UK. Correspondence should be addressed to A.B. ([abradley@sanger.ac.uk](mailto:abradley@sanger.ac.uk)).

Received 6 June 2017; accepted 19 June 2018; published online 16 July 2018; corrected online 31 July 2018; doi:10.1038/nbt.4192



**Figure 1** Frequency of *PigA* loss upon editing with exonic and intronic gRNAs in mouse ES cells. **(a)** Experimental design. Cells were transfected with separate *PiggyBac* transposons carrying gRNA and *Cas9* genes and selected for stable transposition. *PigA*-negative cells (green) were sorted, single-cell clones isolated, the region around the cut site amplified, sequenced and mapped to the reference genome. **(b)** Examples of *PigA* editing revealed by FLAER staining, for two gRNAs and one control. Numbers on the *x*-axis identify individual gRNAs (**Supplementary Table 1**). **(c)** Frequency of *PigA* loss caused by *Cas9* with intronic and exonic gRNAs (**Supplementary Table 1**;  $N = 6$  biologically independent cell cultures). Each circle represents one cell culture. NC: negative control, a guide targeting *Cd9*. Thick bars represent exons, hollow ones indicate UTRs.

To understand what genetic changes underlie the generation of *PigA*-deficient cells, we amplified a 5.7-kb region around exon 2 from pools of cells edited with three selected gRNAs introduced by *PiggyBac* transposition, and sequenced the PCR products using the PacBio platform. We observed a depletion of read coverage on a kilobase-scale around the cut sites, consistent with the presence of large deletions (**Fig. 2a**). Cells edited with intronic guides and sorted for loss of *PigA* expression generally exhibited loss of the adjacent exon. If intronic regulatory sequences were present around the exon, the DNA of cells sorted for retention of *PigA* expression would be wild type or contain small indels around the cut site. However, the most frequent lesions in these cells were deletions extending many kilobases up- or downstream, away from the exon. We conclude that, in most cases, loss of *PigA* expression was likely caused by loss of the exon, rather than damage to intronic regulatory elements.

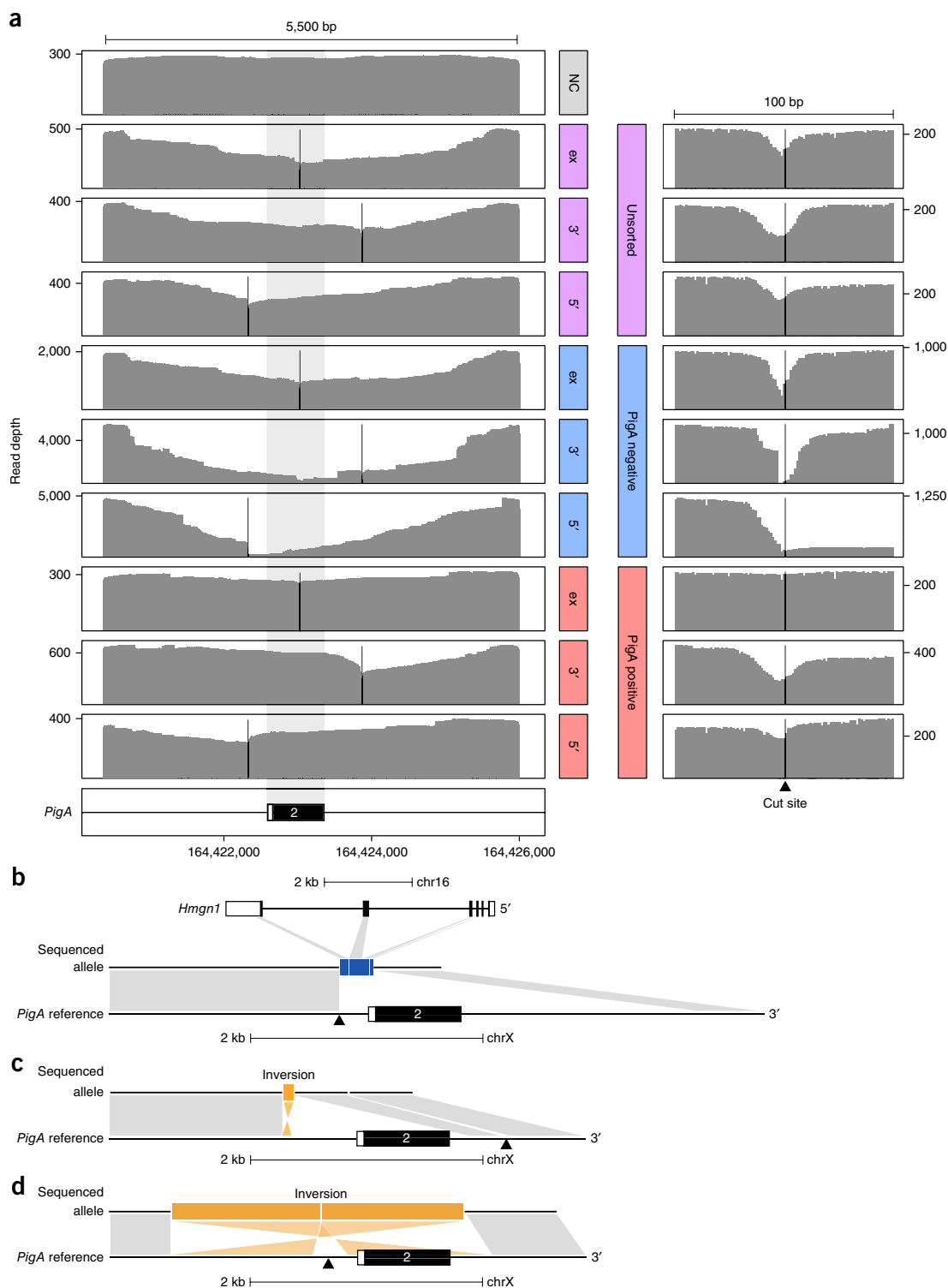
Clustering of PacBio reads yielded 183 unique, edited, high-quality alleles derived from three different gRNAs. These alleles ranged from simple deletions and insertions to complex rearrangements (**Fig. 2a,b**, **Supplementary Table 2** and **Supplementary Data 1**). One of the alleles contained an insertion with a perfect match to four consecutive exons derived from the *Hmgn1* gene (**Fig. 2b**). We speculate this represents a *de novo* insertion from the spliced and reverse-transcribed RNA, rather than from one of the pseudogenized forms of *Hmgn1*, as the pseudogenes diverge in sequence from the observed insertion.

To fully characterize a variety of edited *PigA* loci, we isolated single-cell clones. The *PigA* loci around the gRNA target site were amplified

using PCR primer pairs positioned progressively further apart (up to 16 kb), until amplicons were generated. These were sequenced using conventional Sanger sequencing technology (**Supplementary Fig. 2a**). This strategy allowed us to recover an allele in most cases (133/141, 94%; **Supplementary Tables 2** and **3** and **Supplementary Data 1**).

Simple deletions overlapping both the cut site and the exon were found in almost three-quarters (69/93) of *PigA*-deficient alleles generated by single, intronic gRNAs (**Supplementary Fig. 2b,c**). The deletions varied in size, the largest spanning 9.5 kb. The remaining events were deletions combined with large insertions or more complex, multiple-lesion alleles. We obtained similar results using electroporation of RNP (**Supplementary Fig. 1b**). To assess the frequency of large deletions without strong selection for that outcome, we used an exonic gRNA causing 97% *PigA* loss. Although two-thirds of alleles (32/48) from *PigA*-deficient cells had indels <50 bp, as expected, >20% (10/48) had deletions >250 bp, extending up to 6 kb (**Supplementary Fig. 2d**). Because the deletions generated with the exonic gRNA were bidirectional, this is consistent with the average frequency of generating *PigA*-deficient cells with intronic guides positioned 263–520 bp from an exon (~12%).

Notably, 23 of 133 recovered alleles contained additional lesions (single-nucleotide polymorphisms (SNPs), indels, large deletions and insertions) that were non-contiguous with the lesion at the cut site. In 13 out of 23 cases, the only exonic lesion detected was non-contiguous with the cut site (**Fig. 2c**). Furthermore, we observed alleles in which the intronic gRNA caused an inversion of a region containing the exon

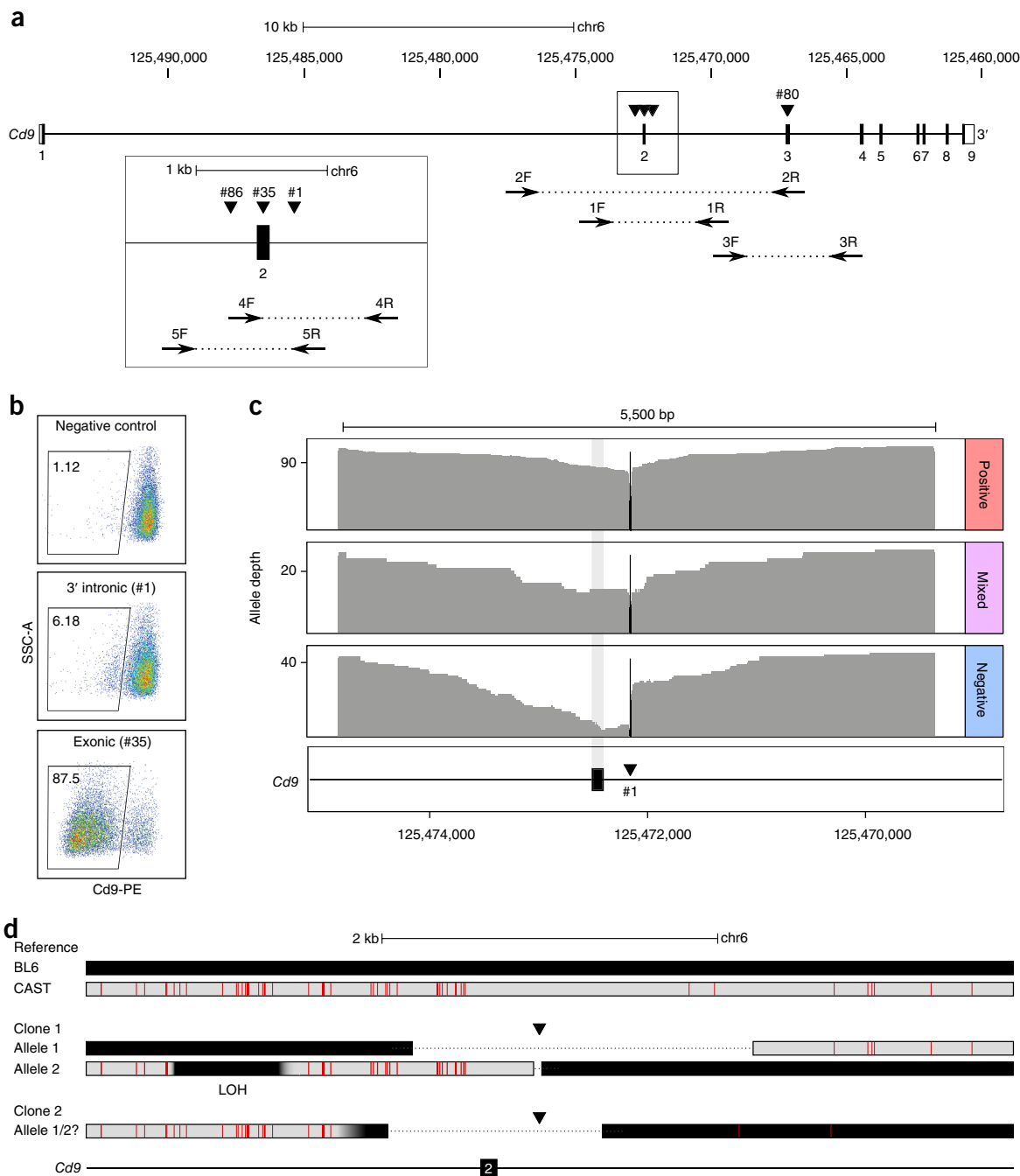


**Figure 2** Analysis of the *PigA* locus edited with selected gRNAs. **(a)** Coverage of PacBio reads at the *PigA* locus. The locus was PCR-amplified from a pool of cells sorted for *PigA* expression (or from the unsorted population), and the resulting products were sequenced using the PacBio platform. The right panel depicts a 100-bp region centered at the cut site. NC: negative-control gRNA, ex: exonic gRNA (#56), 5': 5' intronic gRNA (#15), 3': 3' intronic gRNA (#10). The cut site of the gRNA (between 3<sup>rd</sup> and 4<sup>th</sup> nucleotide from the PAM sequence) is indicated with a vertical black bar. Genomic position is given with respect to the GRCm38 reference genome.  $N = 1$ . **(b–d)** Examples of alleles. The bottom diagram line of each panel represents the *PigA* reference allele around exon 2, the diagram line immediately above shows the structure of the sequenced allele. **(b)** The top diagram line shows the genomic *Hmgn1* gene structure; note the scale differs from that of *PigA* gene. **(c)** Exonic lesion non-contiguous with the cut site. **(d)** Inversion of a region containing the exon. Black horizontal line: direct reference match; orange bar: inversion; blue bar: insertion from another part of the genome; black arrowhead: gRNA target site. Gray and orange shadows represent, respectively, direct and inverted match between the reference and the sequenced allele. Lack of shadow at the reference locus represents a deletion in the sequenced allele.

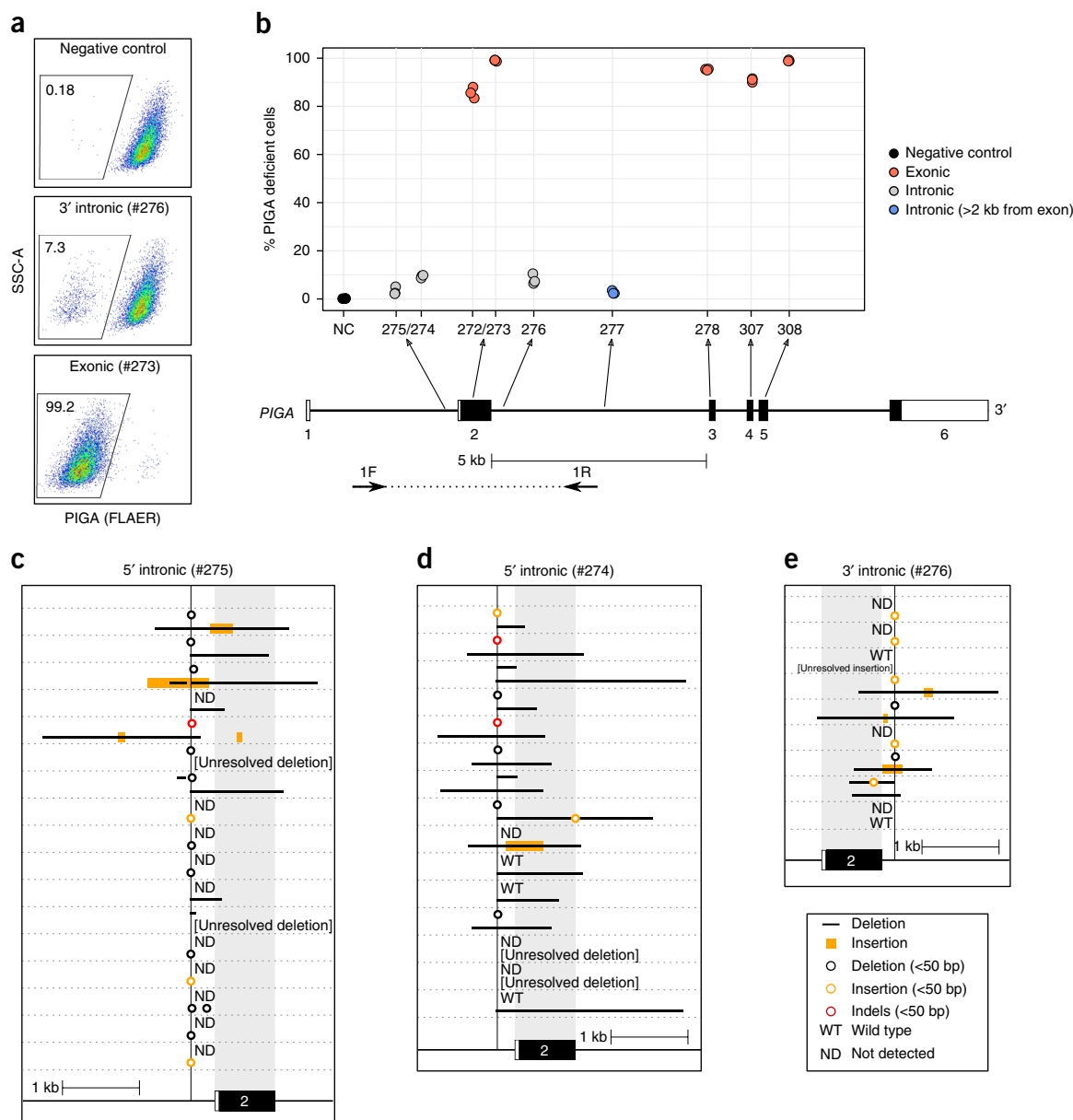
(Fig. 2d). Had the assessment been limited to the immediate vicinity of the cleavage site, such alleles would have been misclassified as wild type, and their phenotypic consequences would have been underestimated.

Insertions were present in 35 out of 133 recovered alleles. We could not find convincing local mapping for insertions shorter than 7 bp (13 alleles), which we speculate to be mostly non-templated nucleotides.

The large majority of other insertions were constituted from sequence, which mapped to the *PigA* locus and encompassed inversions and duplications ranging from 11 bp to 2.5 kb (17 alleles; Fig. 2c,d and Supplementary Fig. 2c). The remaining five alleles contained DNA sequences that mapped to other parts of the mouse genome, such as interspersed repeats, or to exogenous, transfected sequences.



**Figure 3** Analysis of Cas9 editing at the autosomal *Cd9* locus in mouse ES cells. Experimental setup is analogous to the *PigA* experiment in Figure 1a. A mouse ES cell line derived from an F1 cross between CAST and BL6 was used. (a) Positions of primer pairs and gRNAs (Supplementary Tables 1 and 6). Genomic position is given with respect to the GRCh38 reference genome. (b) Examples of *Cd9* editing revealed by antibody staining, for two gRNAs and one control (Supplementary Table 1;  $N = 7$  biologically independent cell cultures). (c) PacBio alleles derived from *Cd9*-positive, mixed (bimodal) and *Cd9*-negative, individually sequenced single-cell clones, displayed as a pileup. Display conventions as in Figure 2.  $N = 1$ . (d) Recombinant alleles. Two of the sequenced single-cell clones contained alleles indicative of a cross-over event between the homologous chromosomes. Red vertical bars in CAST allele (gray bar) indicate positions of sequence divergence from the BL6 reference genome (black bar), dotted black line indicates missing sequence (deletion), thin black line indicates an intron. LOH: loss of heterozygosity.



**Figure 4** Frequency of *PIGA* loss upon editing with exonic and intronic gRNAs and structure of the recovered alleles in human RPE1 cells. Cas9-expressing cells were transfected with *PiggyBac* transposons carrying a gRNA and selected for stable transposition. *PIGA*-negative cells were sorted, single-cell clones isolated, and the region around the cut site amplified, sequenced and mapped to the reference genome. **(a)** Examples of *PIGA* editing revealed by FLAER staining, for two gRNAs and one control. **(b)** Frequency of *PIGA* loss caused by Cas9 with intronic and exonic gRNAs (**Supplementary Table 1**;  $N = 3$  biologically independent cell cultures). Position of the primers with the largest span (6 kb) is indicated. **(c–e)** Recovered alleles. 5' intronic guide #275 **(c)**, 5' intronic guide #274 **(d)**, 3' intronic guide #276 **(e)**. The position of the gRNA is shown as a vertical line intersecting with the *PIGA* gene structure. Pure insertions and deletions of <50 bp are indicated with orange and black circles, respectively. Combined insertion/deletion events of <50 bp and SNPs ('indels') are indicated with a red circle. Black lines represent deletions >50 bp. Orange bars indicate size of the >50-bp insertions (but not their map position). Thin, horizontal, dashed line separates clones.

Six alleles did not contain lesions overlapping the nearest exon. Three of these were also wild type around the cut sites and are likely to contain lesions in other exons or larger rearrangements. The remaining three alleles contained only intronic lesions, which may interfere with splicing. In eight cases, it was not possible to recover any product with exon-spanning primers (**Supplementary Fig. 3a**, black primer pairs). To understand this class of events, we performed additional PCRs targeting each end of the *PigA* locus (**Supplementary Fig. 3a**,

gray primer pairs). In three cases, just one end or neither end of the locus could be amplified, suggesting a larger deletion. In the remaining five cases, both ends were amplified. Since no product connecting the two ends could be obtained, these are likely to be translocations, inversions or large insertions (**Supplementary Table 4**).

To understand the diversity of potential deletion outcomes, we have repeated our original experiment in biological quadruplicate using the 5' intronic gRNA. Cells with large deletions were enriched by



sorting for *PigA*-negative cells and deletion fingerprints were generated by PCR. Each biological replicate differed substantially, despite a large number of unique deletion events sampled, indicating that the diversity of potential deletion outcomes is vast (**Supplementary Fig. 4** and **Supplementary Note**).

Given that *PigA* is mono-allelic in the XY ES cells used in this study we wished to exclude the possibility that the observations reflect some peculiarity of the lack of a homolog. The autosomal *Cd9* locus was selected for this purpose as it is non-essential in ES cells and its protein product can be readily detected by cell surface staining. An exonic guide yielded 88% *Cd9* loss, while 5' and 3' intronic guides generated 4.2% and 5.4% *Cd9* loss, respectively (**Fig. 3a,b** and **Supplementary Table 1a**). Taking into account a 1.6% background of *Cd9*<sup>low</sup> cells in the untransfected condition, we estimate the true proportion of *Cd9* loss due to intronic cutting to be between 2.6–3.8%. This is consistent with results at the *PigA* locus, assuming both *Cd9* alleles have to be destroyed to prevent *Cd9* expression.

To describe the genetic events underlying *Cd9* loss, we isolated single-cell clones edited with the 3' intronic guide, ascertained their expression status by flow cytometry and sequenced the area around the cut site using PacBio and Sanger technologies. The largest deletion spanned 5.5 kb. A pileup of 185 resolved alleles derived from 93 single-cell clones shows a clear enrichment for deletions overlapping the exon in clones negative for *Cd9* compared to positive clones and ones exhibiting a mixture of *Cd9*-positive and *Cd9*-negative cells (**Fig. 3c**). The bimodal expression pattern of some of the clones may be the result of a mixed clone or a protracted repair event that was resolved during clone outgrowth. The haplosufficient nature of the *Cd9* gene is demonstrated by the fact that we could detect at least one allele with an intact exon in all but one of the 66 *Cd9*-positive and mixed clones. Similarly, only one of the 27 *Cd9* negative clones had an intact exon, this exception presumably harboring other undetected lesions. We have further confirmed by PCR genotyping that large deletions are a common outcome in single-cell clones edited at the *Cd9* locus using additional intronic and exonic guides (**Supplementary Table 5** and **Supplementary Note**).

The experiment at the *Cd9* locus was performed in mouse ES cells derived from an F1 cross between *Mus musculus* (BL6) and *Mus musculus castaneus* (CAST) mouse strains, which allowed us to distinguish the homologous chromosomes. In no case was the repair outcome identical between homologs within a clone, despite 15 alleles reoccurring between clones. This result is consistent with the great diversity of outcomes at the *PigA* locus. Just over half of the edited clones (52 out of 93) contained precisely one CAST and one BL6 allele, as expected. Notably, in 18 clones only one allele was detected, potentially due to translocations, very large deletions, insertions or inversions, monosomy or loss of heterozygosity (LOH) either local or chromosome-wide. 21 clones contained an abnormal number of alleles, which could have resulted from a mixed clone, large duplication, repair events happening during clone outgrowth or aneuploidy induced by Cas9 cutting. Finally, two clones contained recombinant BL6-CAST alleles (**Fig. 3d**). In one case, an LOH event distal to the breakpoints converted part of the CAST allele to BL6. In another case, the BL6-CAST crossover boundary did not coincide with the breakpoint. We conclude that the creation of these alleles likely involved interhomolog strand invasion as they cannot be explained by a simple rejoining of the resected ends of two broken chromosomes.

To investigate whether the observed on-target extensive DNA repair-associated damage is an intrinsic property of undifferentiated mouse ES cells, we examined the consequence of editing in a human differentiated cell line. An immortalized human female retinal pigment epithelial

cell line (RPE1) was used. Although this is a female cell line, X-inactivation renders it functionally hemizygous at the *PIGA* locus. Editing *PIGA* with single exonic and intronic gRNAs delivered with *PiggyBac* vectors, resulted in a loss of *PIGA* at frequencies comparable to those observed in mouse ES cells (**Fig. 4a,b**). PCR genotyping and Sanger sequencing of 41 *PIGA*-deficient single-cell clones edited with intronic gRNAs revealed large deletions, insertions, inversions and non-contiguous lesions overlapping the exon (**Fig. 4c–e**). In some clones only one small, intronic indel allele was detected, which we interpret as an inconsequential edit of the inactive chromosome, coupled with a loss-of-function lesion on the active X-chromosome; the lesion would inactivate one or both primer binding sites.

Similar results were obtained in lineage-negative cells from the bone marrow of mice homozygous for a *Cas9-GFP* cassette at the *Rosa26* locus. Progenitor cells enriched by removal of differentiated cells on magnetic columns were electroporated with a crRNA:trRNA complex against the GFP locus, and GFP-negative single-cell clones were isolated and genotyped around the cut site with three different primer pairs spanning up to 3.6 kb. At least one large deletion product between 100 bp and ~3 kb in size was detected in 35 out of 96 clones (**Supplementary Fig. 5a,b**). We verified eight deletion products by Sanger sequencing across the deletion junction (**Supplementary Fig. 6a**). Only wild-type-size products were detected in the remaining clones and none of the 96 control clones exhibited any deletion bands (**Supplementary Table 5**, “progenitor” experiment).

The editing in this study was conducted at actively transcribed loci in normal ES cells and progenitor cells, both with intact DNA repair processes, as well as in an immortalized, differentiated human cell line; each are surrogates for various clinical editing applications. We show that extensive on-target genomic damage is a common outcome at all loci and in all cell lines tested. Moreover, the genetic consequences observed are not limited to the target locus, as events such as loss-of-heterozygosity will uncover recessive alleles, whereas translocations, inversions and deletions will elicit long-range transcriptional consequences. Given that a target locus would presumably be transcriptionally active, mutations that juxtapose this to one of the hundreds of cancer-driver genes may initiate neoplasia. In the clinical context of editing many billions of cells, the multitude of different mutations generated makes it likely that one or more edited cells in each protocol would be endowed with an important pathogenic lesion. Such lesions may constitute a first carcinogenic ‘hit’ in stem cells and progenitors, which have a long replicative lifespan and may become neoplastic with time. Such a circumstance would be similar to the activation of *LMO2* by pro-viral insertion in some of the early gene-therapy trials, which caused cancer in these patients<sup>30</sup>. Results reported here also illustrate a need to thoroughly examine the genome when editing is conducted *ex vivo*. As genetic damage is frequent, extensive and undetectable by the short-range PCR assays that are commonly used, comprehensive genomic analysis is warranted to identify cells with normal genomes before patient administration.

## METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGMENTS

We wish to thank M. Friedrich for sharing his gRNA expression construct and technical advice, E. Metzakopian for technical advice and critical reading of the

manuscript, G. Rutledge for critical reading of the early manuscript, A. Ferguson-Smith for the CAST/B6 hybrid ES cells, P. Liu and X. Gao for mCherry/GFP reporter cells, S. Jackson's group for the Cas9-expressing RPE1 cell line and the Cytometry Core Facility for assistance with cell sorting. This work was supported by the Wellcome Trust Grant number 098051.

#### AUTHOR CONTRIBUTIONS

M.K. performed most of the experiments and analyzed the data. K.T. performed the primary cell work. A.B. supervised the project. All authors contributed to writing of the manuscript.

#### COMPETING INTERESTS

The authors declare no competing interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

- Cornu, T.I., Mussolino, C. & Cathomen, T. Refining strategies to translate genome editing to the clinic. *Nat. Med.* **23**, 415–423 (2017).
- Kim, S., Kim, D., Cho, S.W., Kim, J. & Kim, J.-S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **24**, 1012–1019 (2014).
- 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).
- Frock, R.L. *et al.* Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat. Biotechnol.* **33**, 179–186 (2015).
- Xie, F. *et al.* Seamless gene correction of  $\beta$ -thalassaemia mutations in patient-specific iPSCs using CRISPR/Cas9 and piggyBac. *Genome Res.* **24**, 1526–1533 (2014).
- Guilinger, J.P., Thompson, D.B. & Liu, D.R. Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat. Biotechnol.* **32**, 577–582 (2014).
- Kleinstiver, B.P. *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
- Ran, F.A. *et al.* Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).
- Slaymaker, I.M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).
- Tsai, S.Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* **32**, 569–576 (2014).
- Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
- Tsai, S.Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
- Koike-Yusa, H., Li, Y., Tan, E.-P. & Velasco-Herrera, M.D.C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* **32**, 267–273 (2014).
- van Overbeek, M. *et al.* DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks. *Mol. Cell* **63**, 633–646 (2016).
- Tan, E.P., Li, Y., Velasco-Herrera, M.D.C., Yusa, K. & Bradley, A. Off-target assessment of CRISPR-Cas9 guiding RNAs in human iPS and mouse ES cells. *Genesis* **53**, 225–236 (2015).
- Hendel, A. *et al.* Chemically modified guide RNAs enhance CRISPR-Cas genome editing in human primary cells. *Nat. Biotechnol.* **33**, 985–989 (2015).
- Ghezraoui, H. *et al.* Chromosomal translocations in human cells are generated by canonical nonhomologous end-joining. *Mol. Cell* **55**, 829–842 (2014).
- Weinstock, D.M., Elliott, B. & Jasin, M. A model of oncogenic rearrangements: differences between chromosomal translocation mechanisms and simple double-strand break repair. *Blood* **107**, 777–780 (2006).
- Canver, M.C. *et al.* Characterization of genomic deletion efficiency mediated by CRISPR/Cas9 in mammalian cells. *J. Biol. Chem.* **289**, 21312–21324 (2014).
- Kraft, K. *et al.* Deletions, inversions, duplications: engineering of structural variants using CRISPR/Cas in mice. *Cell Rep.* **10**, 833–839 (2015).
- Boroviak, K., Doe, B., Banerjee, R., Yang, F. & Bradley, A. Chromosome engineering in zygotes with CRISPR/Cas9. *Genesis* **54**, 78–85 (2016).
- Boroviak, K., Fu, B., Yang, F., Doe, B. & Bradley, A. Revealing hidden complexities of genomic rearrangements generated with Cas9. *Sci. Rep.* **7**, 12867 (2017).
- Parikh, B.A., Beckman, D.L., Patel, S.J., White, J.M. & Yokoyama, W.M. Detailed phenotypic and molecular analyses of genetically modified mice generated by CRISPR-Cas9-mediated editing. *PLoS One* **10**, e0116484 (2015).
- Shin, H.Y. *et al.* CRISPR/Cas9 targeting events cause complex deletions and insertions at 17 sites in the mouse genome. *Nat. Commun.* **8**, 15464 (2017).
- Gasparini, M. *et al.* CRISPR/Cas9-mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. *Am. J. Hum. Genet.* **101**, 192–205 (2017).
- Roberts, S.A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
- Sinha, S. *et al.* Microhomology-mediated end joining induces hypermutagenesis at breakpoint junctions. *PLoS Genet.* **13**, e1006714 (2017).
- Yang, Y., Sterling, J., Storici, F., Resnick, M.A. & Gordenin, D.A. Hypermutability of damaged single-strand DNA formed at double-strand breaks and uncapped telomeres in yeast *Saccharomyces cerevisiae*. *PLoS Genet.* **4**, e1000264 (2008).
- Tichy, E.D. *et al.* Mouse embryonic stem cells, but not somatic cells, predominantly use homologous recombination to repair double-strand DNA breaks. *Stem Cells Dev.* **19**, 1699–1711 (2010).
- Hacein-Bey-Abina, S. *et al.* A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N. Engl. J. Med.* **348**, 255–256 (2003).

## ONLINE METHODS

**Mouse ES cell culture and transfection.** gRNA-expression vectors contain a U6 promoter with an “F+E” scaffold<sup>31</sup> and a *PGK-Puro-2A-BFP* cassette, flanked by *PiggyBac* repeats. The Cas9-expression vector contains a *Cas9-Blast* cassette expressed from a short EF1 $\alpha$  promoter in a pKLV backbone<sup>13,32</sup>. CAST/BL6 (CB9; a gift from Prof A. Fergusson-Smith), AB2.2 mCherry/GFP reporter (a gift from X. Gao and P. Liu) or JM8.A3 mouse ES cells<sup>33,34</sup> were cultured in M15 media (high-glucose DMEM, with 15% FCS, beta-mercaptoethanol and L-glutamate) on STO-neo-LIF-puro (SNLP) feeder cells.

Complexes of lipofectamine LTX (2.5  $\mu$ l), plus reagent (0.5  $\mu$ l), 200 ng hyperactive *PiggyBac* transposase<sup>35</sup>, 100 ng of the *PiggyBac Cas9-Blast* plasmid and 50 ng of the *PiggyBac gRNA-Puro* plasmid were prepared in 50  $\mu$ l OptiMEM following manufacturer’s instructions. Cells were trypsinized, washed in M15, resuspended in M15+LIF and seeded onto a gelatinized 24-well plate, containing the lipofectamine DNA complexes, at  $3 \times 10^5$  cells per well. From day 2, M15+LIF media containing puromycin (3  $\mu$ g/ml) and blasticidin (10  $\mu$ g/ml) was used. The same setup was used for RPE1 cell line, except the *Cas9-Blast* plasmid was omitted. A similar setup was used for lipofection of RNP complexes with 20 pmol of both hybridized crRNA:trRNA (Sigma) and EnGen Cas9 NLS (NEB). Neon Transfection System (Thermo Fisher Scientific; 1,600 v/10 ms /3 pulses) was used for electroporation of  $1.5 \times 10^5$  cells in buffer R with 6 pmol each of crRNA:trRNA, electroporation enhancer (IDT) and Cas9 protein or 9 pmol each of crRNA:trRNA and Cas9 protein. Around  $3 \times 10^5$  cells were collected on day 14 (or day 17, in case of the RPE1 cells), stained in PBS+0.1% BSA for 30 min at room temperature with 1  $\mu$ g/ml FLAER reagent (Cedarlane) or anti-Cd9-PE antibody (cat. 124805, Biolegend), washed twice and analyzed using a Cytoflex flow cytometer. For single-cell cloning and PacBio experiments, cells were transfected in six-well plates with five times more cells and reagents, expanded onto 10-cm dishes and sorted by fluorescence-activated cell sorting for loss of FLAER or Cd9 staining on day 14 using MoFlow XDP (Beckman Coulter). Single-cell clones were isolated and grown in 96-well plates. DNA was extracted by proteinase K digestion followed by ethanol precipitation. PCR reaction were conducted using primers in **Supplementary Table 6** and LongAMP polymerase (NEB) following manufacturer’s instructions.

**Bioinformatics.** Primers were designed using Primer3-BLAST (**Supplementary Table 6**). Guide RNAs were designed using Benchling and CRISPRscan<sup>36</sup>. Alignment of Sanger-sequenced PCR products was performed using BLAT (v 36) and converted into BAM format using a customized script from T. Marschall (<https://github.com/ALLBio/allbiotc2/tree/master/synthetic-benchmark>). Mixed traces were resolved using PolyPeakParser<sup>37</sup>. Analysis of PacBio data was performed using command line version of SMRT-Link software (pbtranscript 1.0.1.TAG-1470). For *PigA* locus pileup, circular consensus sequences were called with at least one full pass and minimum predicted accuracy of 0.9. Individual *PigA* and *Cd9* alleles were reconstructed by following “Running Iso Seq using SMRTLink” tutorial on github, except “\_targeted\_isoseq” option was used at the clustering step. Resulting alleles were mapped to the reference genome using bwa mem (v 0.7.17-r1188). In case of

the *PigA* locus, mapped reads were clustered further using a custom script. Genome coverage was calculated with “bedtools genomecov -dz” (v 2.27.1) using circular consensus sequences (*PigA* locus) or reconstructed alleles (*Cd9* locus). All downstream analysis was performed using custom R (v 3.3.2) and bash scripts and visualized with ggplot2 package. Flow cytometric data were processed with FlowJo (v 10.4.1).

**Mouse bone marrow cell culture and transfection.** Bone marrow cells from a homozygous C57BL/6 CAS9-EGFP knock-in mouse<sup>38</sup> were isolated by flushing tibias and femurs in HBSS (Life Technologies) supplemented with 2% FBS and 10 mM HEPES (Sigma). Lineage negative cells were isolated using Direct Lineage Cell Depletion Kit Mouse (Miltenyi Biotec) and cultured in X-Vivo (Lonza) with 2% FBS, 50 ng/ml stem cell factor, 50 ng/ml thrombopoietin, 10 ng/ml IL-6 (PeproTech). After culturing for 3 h,  $1 \times 10^5$  cells were electroporated (1550 v/20 ms/1 pulse) in buffer T with 44 pmol of preassembled crRNA:trRNA duplex (guide #311, **Supplementary Table 1**; IDT) using the Neon Transfection System. GFP-negative cells were sorted 4 d after the electroporation and plated into Methocult M3434 media (6,000 cells per 3 ml, StemCell Technologies). Seven days later, single colonies were picked into 25  $\mu$ l of direct PCR lysis buffer (Qiagen) with 1  $\mu$ g/ml proteinase K and analyzed by PCR (**Supplementary Fig. 5** and **Supplementary Table 5**).

**Life Sciences Reporting Summary.** Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

**Data availability.** PacBio sequencing data are accessible at the European Nucleotide Archive under accession numbers ERS2396492 (*PigA*) and ERS2396493 (*Cd9*). Barcoding information is in **Supplementary Data 2**. Correspondence and requests for materials, additional data and code should be addressed to A.B. ([abradley@sanger.ac.uk](mailto:abradley@sanger.ac.uk)).

- Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).
- Hsu, P.D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
- Strogantsev, R. *et al.* Allele-specific binding of ZFP57 in the epigenetic regulation of imprinted and non-imprinted monoallelic expression. *Genome Biol.* **16**, 112 (2015).
- Pettitt, S.J. *et al.* Agouti C57BL/6N embryonic stem cells for mouse genetic resources. *Nat. Methods* **6**, 493–495 (2009).
- Yusa, K., Zhou, L., Li, M.A., Bradley, A. & Craig, N.L. A hyperactive piggyBac transposase for mammalian applications. *Proc. Natl. Acad. Sci. USA* **108**, 1531–1536 (2011).
- Moreno-Mateos, M.A. *et al.* CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* **12**, 982–988 (2015).
- Hill, J.T. *et al.* Poly peak parser: method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products. *Dev. Dyn.* **243**, 1632–1636 (2014).
- Platt, R.J. *et al.* CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell* **159**, 440–455 (2014).



## Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### ▶ Experimental design

#### 1. Sample size

Describe how sample size was determined.

No power calculation was performed. The relevant effect sizes in question are on the order of magnitude higher than the variability of the assay, so no formal statistical tests were deemed necessary to support the points made in the manuscript

#### 2. Data exclusions

Describe any data exclusions.

No data was excluded from the manuscript

#### 3. Replication

Describe whether the experimental findings were reliably reproduced.

As described, three to eight independent biological replicates were performed with minimal variation and all were successful.

#### 4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

Samples (cell lines) were assigned to all experimental groups as they are not a limited resource, so no randomization was necessary.

#### 5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

While blinding could theoretically be applied to the experimental setup, the effect sizes were on the order of magnitude higher than the variability of the assay, which combined with the simple nature of the readouts (DNA sequences, single-stain flow cytometry) left little room for conclusion-altering unconscious manipulation.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

#### 6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

- |                                     |   |
|-------------------------------------|---|
| n/a                                 | Confirmed   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> The <u>exact sample size</u> ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)                         |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly                              |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A statement indicating how many times each experiment was replicated  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of any assumptions or corrections, such as an adjustment for multiple comparisons  |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> The test results (e.g. $P$ values) given as exact values whenever possible and with confidence intervals noted   |
| <input type="checkbox"/>            | <input checked="" type="checkbox"/> A clear description of statistics including <u>central tendency</u> (e.g. median, mean) and <u>variation</u> (e.g. standard deviation, interquartile range)                               |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clearly defined error bars   |

See the web collection on [statistics for biologists](#) for further resources and guidance.

## ► Software

Policy information about [availability of computer code](#)

### 7. Software

Describe the software used to analyze the data in this study.

R 3.3.2, BLAT 36, pbtranscript 1.0.1.TAG-1470, FlowJo 10.4.1, bedtools 2.27.1, samtools 1.6, bwa mem 0.7.17-r1188

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* [guidance for providing algorithms and software for publication](#) provides further information on this topic.

## ► Materials and reagents

Policy information about [availability of materials](#)

### 8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

There is no such restrictions.

### 9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

We performed CRISPR/Cas9 knock-out experiments, followed by genetic characterization of clones sorted for expression-level of targets in question, which constitutes a validation in itself. FLAER reagent used in the study, which is a fusion of protoaerolysin and FITC fluorophore, was obtained from Cedarlane and is certified for in-vitro diagnostic use (<https://www.cedarlanelabs.com/Products/Detail/FL2S-C?lob=AllProducts>). It has been extensively tested by our collaborators, e.g. Koike-Yusa, H., et al. *Nat. Biotechnol.* 2014..

### 10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

CAST/BL6 CB9 cells were a gift from Prof A. Fergusson-Smith, AB2.2 mCherry/GFP reporter were a gift from Dr. Xiufei Gao and Prof. P. Liu, RPE1-p53-neg-Cas9+ cell line was a gift from S.Jackson's group and JM8.A3 mouse embryonic stem cells were produced by Steve Pettitt in this group.

b. Describe the method of cell line authentication used.

A low passage cells were used, which are routinely employed for gene targeting and embryo injection, guaranteeing their embryonic stem cell status.

c. Report whether the cell lines were tested for mycoplasma contamination.

All cell lines have tested negative for mycoplasma.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

No such cell lines were used.

## ► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

### 11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used.

Policy information about [studies involving human research participants](#)

### 12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

Humans were not subjects of this study.

---

## Erratum: Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements

Michael Kosicki, Kärt Tomberg & Allan Bradley

*Nat. Biotechnol.* doi:10.1038/nbt.4192; corrected online 31 July 2018

In the version of this article initially published online, four figure citations were incorrect on p.2: left-hand column, after “complex rearrangements,” “Supplementary Fig. 2a,b” should have been “Fig. 2a,b”; right-hand column, in three places, the citation for “Supplementary Fig. 3...” should have been for “Supplementary Fig. 2.” The errors have been corrected for the print, PDF and HTML versions of this article.