nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-30480-8 OPEN

Cas9-induced large deletions and small indels are controlled in a convergent fashion

Michael Kosicki[1], Felicity Allen[2], Frances Steward[3], Kärt Tomberg[3], Yangyang Pan[3] & Allan Bradley[3]✉

Repair of Cas9-induced double-stranded breaks results primarily in formation of small insertions and deletions (indels), but can also cause potentially harmful large deletions. While mechanisms leading to the creation of small indels are relatively well understood, very little is known about the origins of large deletions. Using a library of clonal NGS-validated mouse embryonic stem cells deficient for 32 DNA repair genes, we have shown that large deletion frequency increases in cells impaired for non-homologous end joining and decreases in cells deficient for the central resection gene *Nbn* and the microhomology-mediated end joining gene *Polq*. Across deficient clones, increase in large deletion frequency was closely correlated with the increase in the extent of microhomology and the size of small indels, implying a continuity of repair processes across different genomic scales. Furthermore, by targeting diverse genomic sites, we identified examples of repair processes that were highly locus-specific, discovering a role for exonuclease Trex1. Finally, we present evidence that indel sizes increase with the overall efficiency of Cas9 mutagenesis. These findings may have impact on both basic research and clinical use of CRISPR-Cas9, in particular in conjunction with repair pathway modulation.

[1] Lawrence Berkeley National Lab, Berkeley, CA, USA. [2] Wellcome Sanger Institute, Hinxton, UK. [3] The Cambridge Institute of Therapeutic Immunology and Infectious Disease (CITIID), Department of Medicine, University of Cambridge, Cambridge, UK. ✉email: ab2592@cam.ac.uk

The goal of genome engineering is the introduction of a particular genotype to the exclusion of others. A programmable nuclease Cas9 is currently the primary tool of genome engineering in clinical and basic research context. Resolution of the double-stranded break (DSB) induced by Cas9 at a location determined by a guide RNA (gRNA) is the principal cause of Cas9 mutagenesis and templated editing. Since the specific outcome depends primarily on the relative activity of different DNA repair pathways, understanding of their function in genome engineering is crucial.

Cas9 mutagenesis is primarily the result of non-homologous end joining (NHEJ) and microhomology-mediate end joining (MMEJ) repair. NHEJ is initiated by Ku70/Ku80 complex binding to the ends of the break, protecting it from degradation. A cascade of events involving, among others, DNA-PKcs, 53BP1, Xlf, Xrcc4 and Lig4, leads to either a perfect repair or a small indel (<10 bp). Resection of the break by the Mre11-Rad50-Nbs1 (MRN) complex, promoted by Ctip and Brca1, prevents NHEJ. Resected DNA can be repaired through microhomology-mediate end joining (MMEJ), which involves Parp1, Polϑ and ligases Lig1 and Lig3, resulting in larger indels[1].

A frequency spectrum of indels resulting from NHEJ or MMEJ repair in a population of cells, the “indel profile”, is specific to local DNA sequence of the target and generally stable across tested cell lines[2–4]. These indels, usually smaller than 50 bp, can be predicted from DNA sequence with high precision[5–8]. In particular, frequent occurrence of 1 bp insertions templated from around the cut site has been linked to Cas9-induced DSBs with 1 nt 5’ overhang[9,10]. The size of indels is typically increased by NHEJ inhibition, while inhibition of the core MMEJ proteins, such as Parp1 and Polϑ, decreases them[3,11–16]. While predictable and partially malleable, Cas9 mutagenesis often does not lead to the desired genome engineering outcomes.

Cas9 templated editing, which hijacks the homologous recombination (HR) pathway, can lead to well-defined outcomes. If the cell is in S/G2 phase of its cell cycle and if the MRN-initiated resection proceeds further, the DSB can be repaired by HR using either sister chromatid (resulting in perfect repair) or an exogenously provided template. This process involves, among others, Brca2 and Rad51[17]. Templated repair using Cas9 is normally harder to achieve than mutagenesis and therefore a number of studies focused on increasing its frequency. In addition to optimization of transfection conditions and the template itself (e.g. refs. [18,19]), inhibition of NHEJ proteins by pharmacological means is one of the preferred methods (e.g. refs. [20–23]). At least one company plans to test these inhibitors in the context of clinical genome engineering[24]. Some of the alternative strategies include modulation of the cell cycle and of the HR pathway itself[25–29].

While a lot of literature has focused on small indels and templated repair, Cas9 complexed with a single gRNA can also induce large deletions at least kilobases in size and complex lesions, such as translocations, large insertion and non-contiguous lesions at significant frequencies[30–33]. These effects were also noted in conjunction with templated editing in mice[34–37]. Extensive loss-of-heterozygosity by gene conversion and megabase-long deletions were also observed[38–41]. Such outcomes could be pathogenic, and may be hard to genotype. Methods which do not require DSB to introduce templated edits, such as base editing and prime editing, were developed in part to avoid such consequences. Nonetheless, these tools introduce DSBs occasionally, as evidenced by creation of indels, making it likely they can also introduce large deletions[42,43]. Furthermore, it is not well understood, which DNA repair pathways control their creation.

To avoid large deletions and complex lesion, we need to know which repair mechanisms lead to their creation. To study this issue, we have built a library of mouse NGS-validated embryonic stem cells deficient for 32 DNA repair genes, derived from a single clone constitutively expressing Cas9. Using this library, we discovered that NHEJ genes prevent large deletions, while the resection gene *Nbn* and the MMEJ gene *Polq* are necessary for their creation. We also find a strong correlation between the frequency of large deletions and size or microhomology usage of small indels, across a range of deficient clones. This implies that small indels and large deletions are controlled convergently. By targeting multiple genomic sites, we observed some gene deficiencies have highly locus-specific functions and discovered an additional role for exonuclease Trex1. Finally, we have shown that highly efficient Cas9 editing leads to more MMEJ outcomes.

Results

Mouse embryonic stem cell DNA damage repair deficient library. Many DSB repair deficient cell lines have been used to study Cas9 engineering outcomes in the past. However, available lines have often been derived independently, from different mouse strains, tested using different Cas9 vectors and compared to only one or few control clones. This may introduce cryptic variability in such parameters as proliferation rate or Cas9 expression levels, which in turn may confound the effect of DNA repair deficiency. Likewise, recently developed pooled CRISPR drop-out screens are vulnerable to proliferation rate differences that are not related to tested phenotypes. Seeking to avoid these confounders, we built a library of mouse NGS-validated embryonic stem cell deficient for multiple DNA damage repair genes, based on a single clone constitutively expressing Cas9. Mutations in 39 repair genes were introduced using Cas9 complexed with 81 gRNAs. We selected these genes to broadly cover the main DSB repair pathways, NHEJ, MMEJ and HR. We also included a number of exonucleases, expecting some of them might have a role in generating large deletions.

Approximately 800 clones (incl. controls) were created in a single experiment and passaged together, minimizing differences due to handling and reagents batches. Clones with no detectable wild-type allele and no frame-preserving indels at the target site by targeted short-amplicon sequencing were incorporated into the library (with few exceptions, see Methods and Supplementary Data, clones tab). Since genotyping alone does not guarantee full ablation of the protein product, we refer to the clones as NGS-validated. As expected, attempts to mutate a number of HR genes resulted in extensive lethality and an increased number of in-frame indels (see Methods for details). For some of these genes, we have incorporated clones with large in-frame deletions (more than 10 bp), expecting them to be hypomorphic. We have performed Western blot on selected clones, showing that *Lig1* and *Parp1* exhibit clear loss of protein expression, while *Nbn* and *Xlf* clones had no obvious reduction in signal (Supplementary Fig. 1). In light of the phenotypic results, we speculate that either the antibodies for the latter were not detecting the right proteins or that compensatory alternative splicing rescued the protein product levels, but not the function.

In total, we have selected 83 individual clones, with mutations in 32 repair genes. The library also included 12 control clones transfected with non-targeting gRNAs or a gRNA targeting safe harbor *Rosa26* locus (Supplementary Data, clones tab).

Large deletions are prevented by NHEJ and promoted by *Nbn* and *Polq*. Large on-target deletions and complex lesions are a significant and potentially pathogenic outcome of Cas9 mutagenesis, but DNA repair pathways contributing to these outcomes are unknown. To map out these pathways, we have applied a previously developed flow cytometric assay to the arrayed library

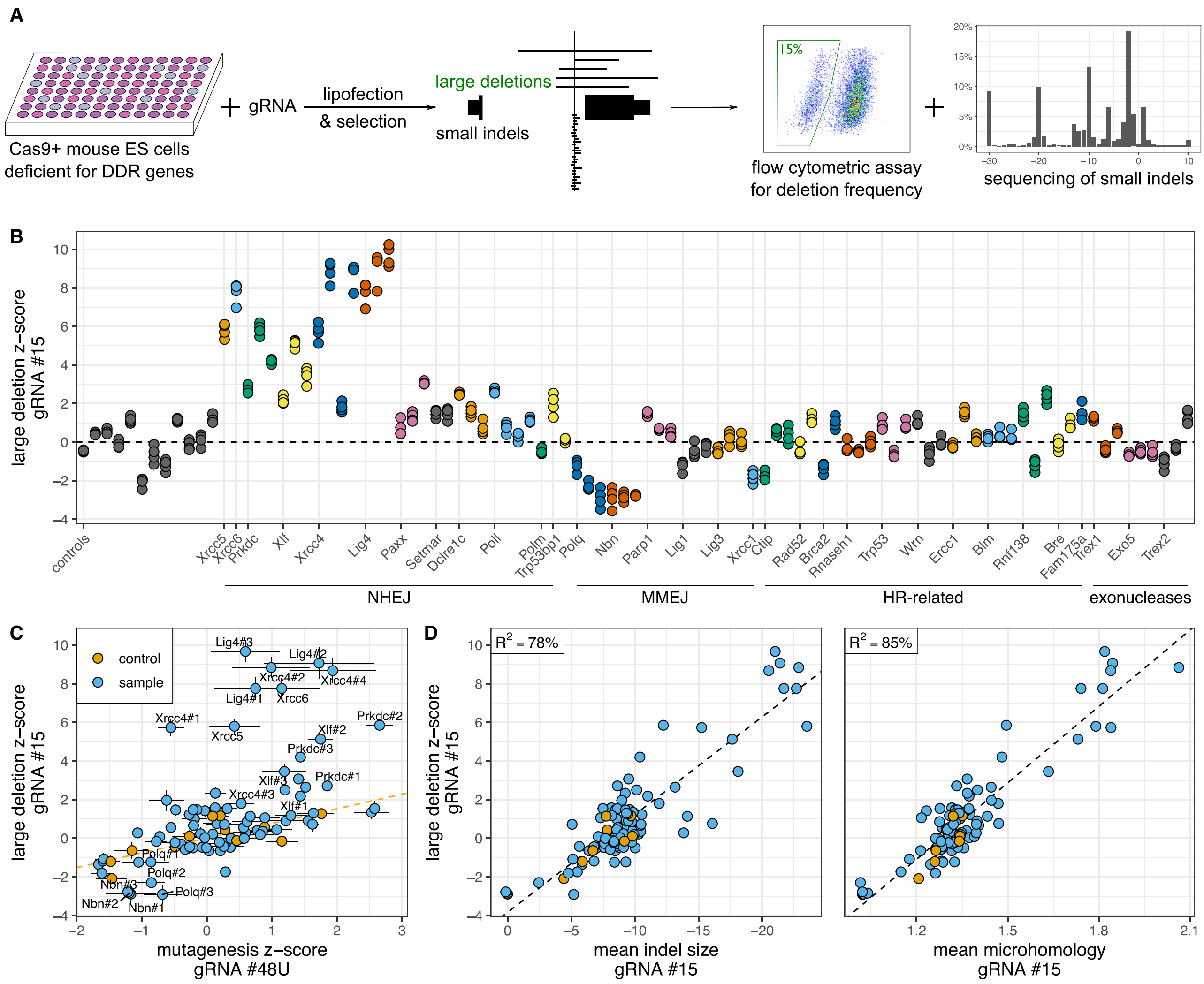

Fig. 1 End joining pathways divergently control the frequency of large deletions caused by Cas9 in mouse ES cells. A Experimental design. Library of Cas9-positive NGS-validated clones deficient for DNA damage repair genes was transfected with individual gRNA-expressing constructs and selected for stable integration. Expression of target genes was measured by flow cytometry, revealing frequency of large deletions (using intronic gRNA #15) or overall mutagenesis (using gRNAs #48U, #48 and #148). Frequency of small indels was established by targeted sequencing of short-range PCR products. **B** Frequency of large deletions caused by Cas9 with intronic gRNA in DNA damage deficient clones, measured by flow cytometry, expressed as a regressed z-score (see Methods). Only the initial clone in a series of clones deficient for the same gene is labeled on the *x* axis. *N* = 4 independent cell cultures. **C** Comparison of large deletion and mutagenesis indices (see Methods). Dashed line indicates best linear fit to control clones (in orange). Error bars are 2xSEM (*N* = 4). **D** Correlation between large deletion z-score (measured by flow cytometry) and the size or microhomology extent of small indels (measured by targeted sequencing). Each dot represents an average readout of an individual clone (*N* = 1–2 biologically independent cell cultures). Negative indel sizes indicate dominance of deletions.

of clonal mouse NGS-validated embryonic stem cell clones deficient for DNA damage repair genes. The assay allows specific detection and isolation of large deletions and complex lesions, as demonstrated by long-read sequencing[31]. We transfected each clone with a gRNA against the intron of the *PigA* gene and measured the frequency of cells that have lost PigA expression (Fig. 1A). As shown before, small indels at this site do not affect PigA expression, and the cells that have lost gene expression harbor large deletions (>260 bp) overlapping the nearest exon or, much more rarely, other complex lesions which explain expression loss (translocations, non-contiguous lesions, insertions containing polyadenylation signals). We will refer to these events collectively as “large deletions”. The fact that in male ES cells there is only one copy of *PigA*, which is located on chromosome X, makes the assay highly sensitive.

We observed a substantial increase in large deletion frequency in clones deficient for the core NHEJ-factors, in particular *Xrcc4*, *Lig4*, *Xrcc5* (Ku80 protein), *Xrcc6* (Ku70), *Prkdc* (DNA-PKcs) and *Xlf* (Fig. 1B). Mutations in other NHEJ genes, such as *Paxx*, *Setmar*, *Dclre1c* (Artemis) and *Poll* did not substantially influence the results, consistent with a minor role they play in this pathway. Conversely, lower frequency of deletions was found in clones mutated at the *Nbn* locus (Nbs1 protein), which is involved in initial resection of DSB leading to MMEJ and HR pathways. Similarly, deletions were less common in clones deficient for *Polq* (Polϑ), a crucial component of MMEJ pathway. Raw frequency of large deletions spanned from almost 30% in *Xrcc4* and *Lig4* deficient clones to around 1% in *Polq* and *Nbn* deficient ones, compared to 12% in control clones and <0.1% background in clones transfected with a non-targeting gRNA against GFP. Since our assay primarily detects deletions spreading in a single direction from the cut site, the true frequency of these lesions is likely 1.5–2 times higher than measured[44].

To control for the expected variability in mutagenic efficiency between individual clones, we compared the deletion frequency with results obtained using exonic gRNA #48U, which tracks mutagenic efficiency. We chose this gRNA as reference, since mutagenesis using other exonic gRNAs was nearing saturation (#48 and #148; Supplementary Fig. 2B, raw frequency). The effect on deletion frequency generally exceeded that on the overall mutagenesis level (Fig. 1C). We conclude that large deletions are prevented by NHEJ repair and promoted by at least some part of MMEJ machinery.

Small indels and large deletions are controlled by the same pathways. The effect of DSB repair pathways on large deletion frequency was qualitatively consistent with the previously described effect of these pathways on local indel profiles[15]. A close quantitative correlation between the two would imply a common mechanism. To see if this is the case, we sequenced a 283 bp area around the cut site of the gRNA #15 we have used to assess large deletion frequency. We found a strong correlation between the average size of microhomology of the sequenced small indels and large deletion frequency as measured by flow cytometry (Pearson $R^2 = 85\%$, Fig. 1D). We also found a moderate, inverse correlation with the average size of the small indels ($R^2 = 78\%$, Fig. 1D). A linear model using both measures was not significantly different from single-measure model employing homology size (p value = 0.13, deviance −1.85, chi-squared test on nested models, residual df = 92, 93). These observations imply a strong commonality of repair mechanisms generating both types of lesions. They also suggest that sequencing of short-range PCR products could be developed as a proxy assay reporting on the changes in frequency of large deletions.

Core end joining genes influence indel profiles of multiple target sites. Screens of DNA damage repair processes often rely on a single locus reporter assay or on composite readouts based on random mutagenesis. However, in vitro biochemical studies show that DNA repair is often highly sequence specific. To distinguish between universal and specific repair processes, we sequenced mutagenized target sites of three gRNAs, each with a distinct indel profile in control clones (Fig. 2A). In particular, gRNA #15 was characterized by preponderance of 1 bp insertions, gRNA #48 by diversity of small indels 1–5 bp in size, while gRNA #148 induced discretely sized deletions (2, 10, 20 bp). We speculate that these profiles reflect relative contribution of NHEJ and MMEJ repair at a given site.

To obtain an overview of relationship between deficient clones, we have calculated Kullback-Leibler divergence between each pair as described in Allen et al.[5] and transformed the resulting divergence matrix using multidimensional scaling (MDS), a nonlinear dimensionality reduction technique similar to principal component analysis (PCA). We found biological replicates to cluster together, indicating good reproducibility (Supplementary Fig. 3A). Furthermore, the majority of clones, including all controls, clustered at the centre of the plot. This indicated most mutants did not influence the indel profile substantially, consistent with the flow cytometry assay. As a further control, we compared the frequency of mutated reads and the frequency of cells which lost expression of the target gene in the flow cytometric assay and found them to match closely for exonic gRNAs #48 and #148 (Supplementary Fig. 3B). As expected, these numbers did not match for the intronic gRNA #15, as in this case the two methods measure mutually exclusive outcomes: the frequency of small indels and the frequency of large deletions.

We asked which deficiencies exhibited similar effects regardless of the target site, and which other deficiencies they clustered with. Mutations in *Xrcc5* and *Xrcc6* genes, whose products form a functional heterodimer (Ku80-Ku70), had very similar, strong effects (Fig. 2B, D). Likewise, indel profiles of *Xrcc4* and *Lig4* mutants clustered together, consistent with the fact Xrcc4 forms a scaffold for Lig4. MMEJ-associated *Polq* and *Nbn* clustered away from NHEJ genes such as *Xrcc*'s 4, 5 and 6 and *Lig4*. As shown previously, NHEJ-deficiencies increased the size of indels, while MMEJ-deficiencies decreased them, although specific role of *Nbn* in indel profile modulation has not been described previously. In general, genes acting earlier in their respective pathways (*Xrcc5*/*Xrcc6* and *Nbn*) had stronger phenotypes than the genes acting later (*Lig4*, Xrcc4 and *Polq*). We note that *Lig1* and *Parp1* clones, despite their knock-outs being confirmed by Western blot (Supplementary Fig. 1A and B), did not have any phenotype in our assay. This is consistent with compensatory action of Lig3 in replication context in case of Lig1[45] and with one of the previous reports on the role of Parp1 obtained in a human cancer cell line HEK293[46,47].

Resection exposes single-stranded DNA, which can participate in repair using microhomology. The extent of microhomology in an indel profile could thus be confounded by the extent of resection. Taking advantage of the wide range of repair outcomes in both control and deficient clones, we decided to investigate the relationship between the two. We found a striking correlation between the average indel size (proxy for resection) and microhomology size for all gRNAs (Pearson R^2 between 71% and 88%, Fig. 2C). On the average, we observed 1 bp more homology for 19–27 bp increase in indel size (depending on gRNA), with the caveat that we do not know if this relationship can be extrapolated beyond the observed intervals. We speculate that this perspective may allow assessment of the relative contribution of deficient genes to resection and microhomology repair, respectively. In particular, we think it is likely that clones close to the regression line (*Trex1*, *Nbn*, *Lig4*, *Xrcc5* and *Xrcc6*) mainly control the extent of resection, while "distal" clones (Polq, some of the significant *Xrcc4* clones, *Dclre1c*, *Prkdc* and *Trp53bp1*) also control the extent of microhomology, at least in some genomic contexts. Consistently, Polϑ, the gene product of the most systematically "regression line-distal" gene, is known to actively generate homologous DNA at the DSB ends[14,48]. We note that clone Ercc1#1 with gRNA #48 was excluded from this particular analysis as a strong outlier, with much larger mean deletion size than controls (−38 bp), without a proportional increase in microhomology usage (1.2 bp). Lack of a replicate sample, relatively low sequencing depth, lack of phenotype for this clone with other gRNAs and of the other two *Ercc1* clones with the same gRNA further solidified our doubts.

Specialized repair pathways affect indel profiles in a locus specific manner. Having focused on indel patterns that were common between the three gRNAs, we turned to gRNAs-specific effects. We found that *Poll* deficiency only had a significant effect on the profile of gRNA #15, *Trex1* on #148 and *Prkdc*, *Dclre1c*, *Trp53bp1*, *Polm* and *Ercc1* on #48 (Fig. 2B, examples of differential indel profiles in Fig. 3A). Clones *Wrn#1* and *Trp53#3* also had specific effect on gRNAs #15 and #48, respectively, but it was far weaker than that of other genes and did not replicate in other independently derived clones. We chose not to explore this further.

We speculated this gRNA specificity is driven by the most prominent indels in each profile. By examining individual indel

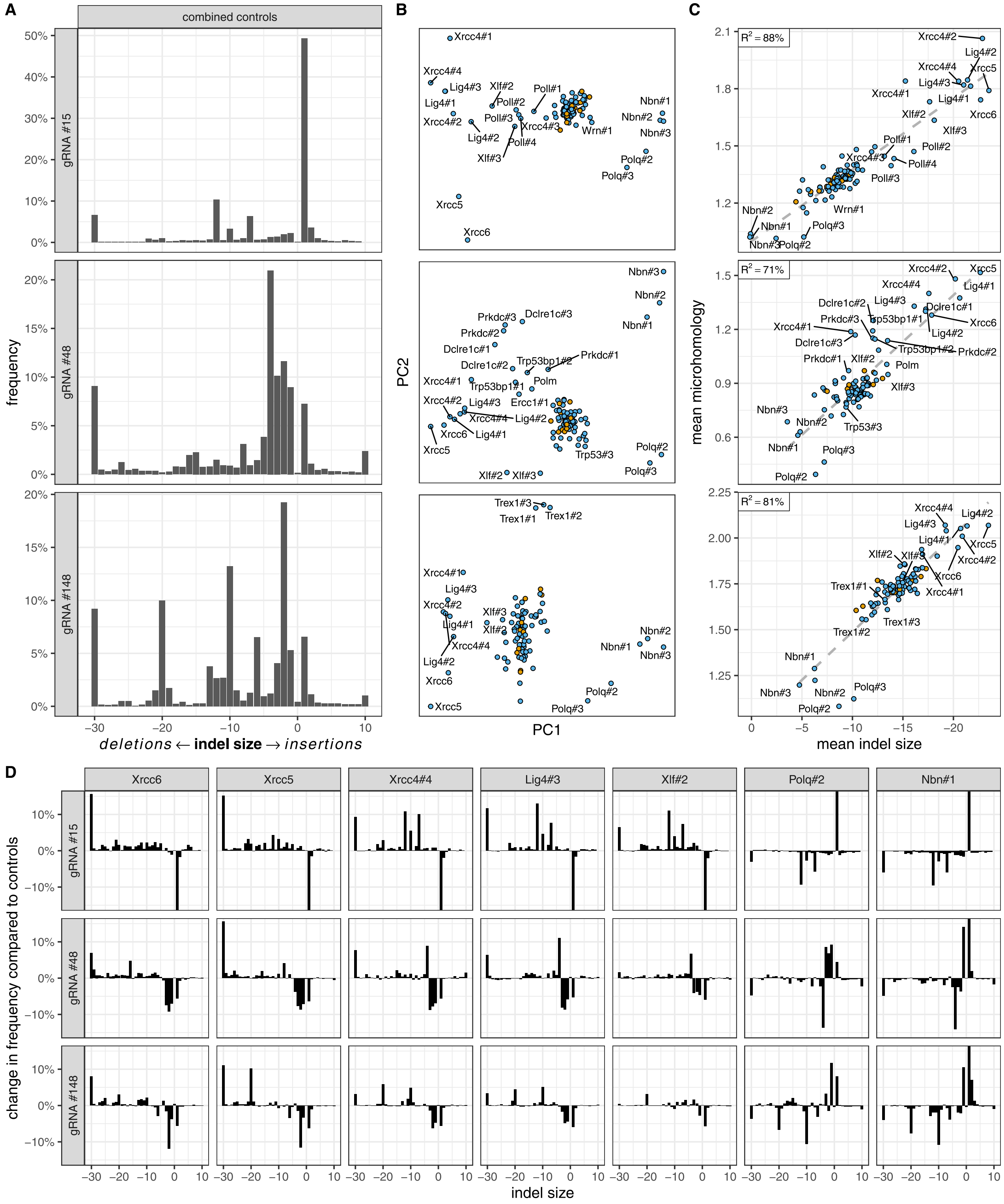

Fig. 2 Core end joining genes influence indel profiles globally. A Indel profiles in combined 12 control samples. **B** Relationships between cell clones based on their indel profiles. Clones significantly different from controls in both replicates are labeled (FDR-corrected $p < 0.01$ from a chi-squared distribution, see Methods). The arrangement of non-significant clones is in Supplementary Fig. 3C. **C** Correlation between mean indel size and microhomology. **D** Relative frequencies of indel sizes compared to controls in deficient clones with a significant impact on all three gRNAs. Indel profiles of other clones with significant impact are in Supplementary Fig. 6. Y axis is truncated at −15% and +15%. In panels (**A**) and (**D**), indel frequencies are aggregated by combined size. Negative numbers represent deletions and positive ones represent insertions. The leftmost and rightmost bars (−30 and 10) combine all larger deletions and insertions, respectively. Biological replicates ($N = 2$) were averaged for clarity. All rows in panels (**B**) and (**C**) relate to the same gRNAs as in panel (**A**). In panels (**B**) and (**C**), controls are in orange and samples are in blue.

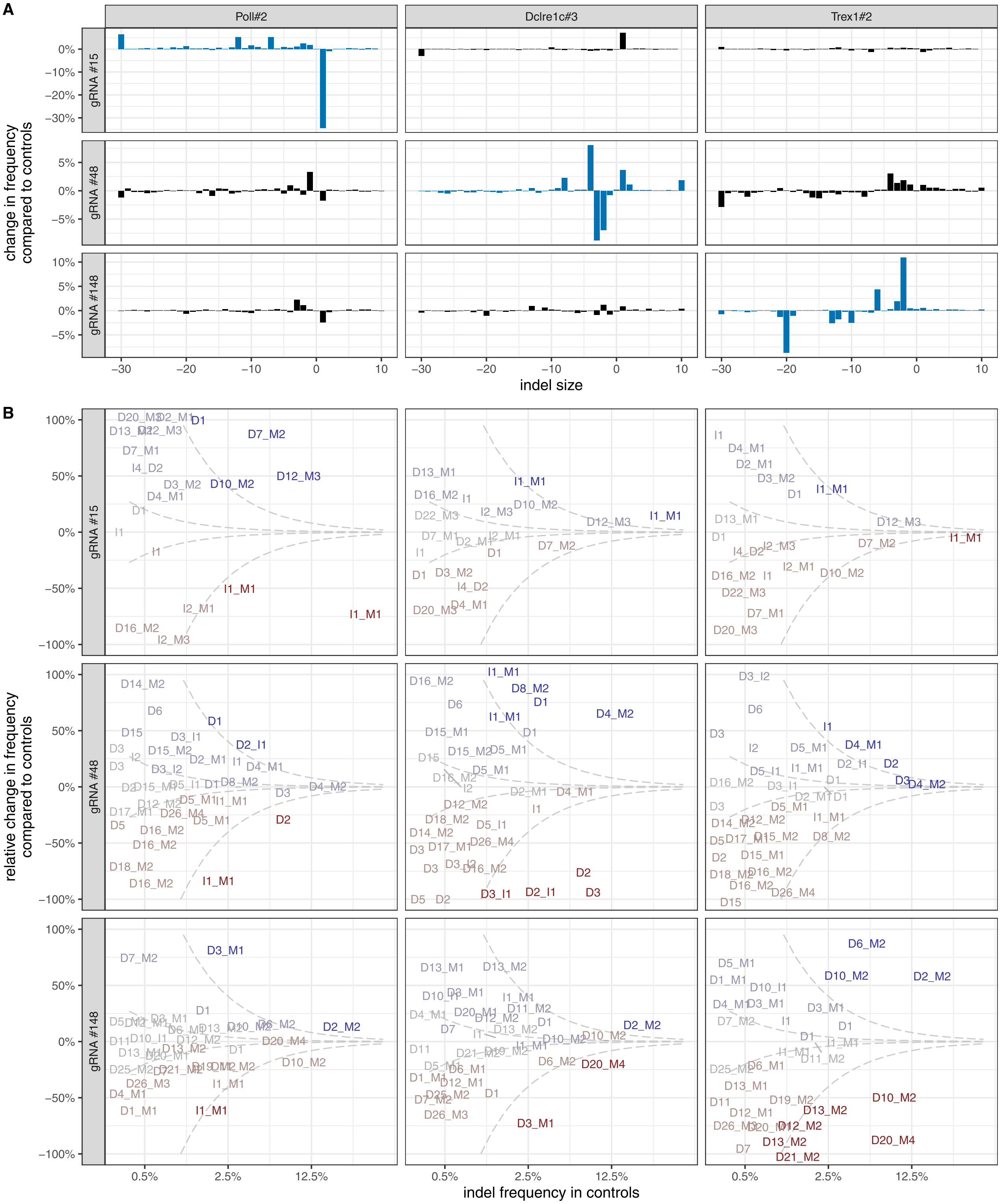

Fig. 3 Deficiencies in specialized DNA damage repair genes influence indel profiles in a locus-specific manner. A Indel profile divergence between controls and selected clones. Blue bars highlight clone/gRNA combinations that were significantly affected (FDR-corrected $p < 0.01$ from a chi-squared distribution, see Methods). **B** Change in frequency of individual indels relative to controls. 'D' = deletion, 'I' = insertion, 'M' = microhomology (see Methods). *X* axis indicates the frequency in control clones. *Y* axis indicates relative change in indel frequency in a given clone relative to control clones. Complete loss of the indel is at −100%, while 100% indicates doubled frequency. The axis is truncated there for display clarity. Only indels present at 0.3% frequency or higher in control clones are shown. Dashed lines indicate absolute change of 0.1% and 1% respectively, color gradations highlight this change. Biological replicates (*N* = 2) were averaged for clarity. All columns relate to the same gRNAs as in panel (**A**).

frequencies we confirmed that 1 bp, microhomology-associated insertions depleted by *Poll* deficiencies in profiles of all tested gRNAs, were most common in the profile of the significantly affected gRNA #15 (Fig. 3B). Analogously, microhomology-containing 7–20 bp deletions prone to *Trex1* depletion were the most prominent outcomes of #148 mutagenesis. Finally, top indels depleted by *Prkdc*, *Dclre1c*, *Trp53bp1*, *Polm* and *Ercc1* were 2–5 bp deletions, commonly induced by gRNA #48 (Dclre1c example in Fig. 3B). The effects of all deficiencies described here are consistent with the literature[49], except *Trex1*, whose function in DSB repair has not been described before.

To learn more about the effect of individual deficiencies, we examined indels that did not conform to the rules broadly laid out above. We found that gRNA #148 induced two different, prominent 10 bp deletions with 2 bp microhomology, whose frequency changed divergently in NGS-validated *Trex1* deficient clones. One of them, the only notable large indel to increase in frequency upon *Trex1* depletion, involved a G-homopolymer. Another divergent indel, a 4 bp deletion with 2 bp microhomology induced by gRNA #48, was promoted by deficiencies in *Prkdc*, *Dclre1c*, *Trp53bp1* and *Polm* (but not *Ercc1*), which otherwise decreased the frequency of 2–5 bp indels. We believe targeting additional loci to find more such apparently divergent outcomes could be useful to learn the rules governing DNA-sequence specific DSB-repair.

Efficient mutagenesis leads to increase in size of small indels. Cas9 has a number of properties that make it likely to interfere with the DSB repair process. Among others, Cas9 can recut the DNA immediately after a perfect repair, may cut both sister chromatids simultaneously, stays bound to DNA after introducing the cut and might possess exonuclease activity. If Cas9 interferes with DSB repair in any of these ways, then manipulating its concentration or activity could result in changed indel profiles. To investigate this issue, we have challenged the library with a low efficiency gRNA #48U, whose target sequence is identical to #48. Unlike #48, #48U's scaffold is expressed as two independent molecules, the crRNA (containing the target-matching sequence) and the tracrRNA. Significantly fewer control cells transfected with this weak gRNA lost PigA expression compared to the strong one (around 12% vs 80%, see Supplementary Fig. 2). We speculate this is a consequence of reduced amount of "productive" gRNA.

To investigate the effect of mutagenic efficiency on repair outcomes, we initially compared the results of the flow cytometric assay using gRNAs of different strengths. Samples transfected using the weak gRNA #48U clustered away not only from #15 samples, which track deletion frequency, but also from the combined cluster of strong exonic gRNAs #48 and #148 (Fig. 4A). This difference was unlikely to be purely driven by the lower flow cytometry read-out with the weak gRNA, because the input for PCA-transformation was mean and standard deviation normalized, which should remove information about the relative magnitude of mutagenesis. Furthermore, #48U samples collected on day 14 post-transfection were further away from the #48 and #148 cluster than samples collected on day 7, which is contrary to the expectation of the observed principal components capturing the magnitude of mutagenic efficiency. We conclude that mutagenic efficiency qualitatively affects the results of the flow cytometry assay.

To test whether mutagenic efficiency affected small indel profiles as well, we compared the sequencing results of gRNAs #48 and #48U. The central cluster of controls and non-affected gene-deficient clones was clearly split between the two gRNAs (Fig. 4B, left). Since mutated alleles were sequenced much more shallowly in #48U samples, which could potentially affect the results, we downsampled all read counts to the lowest common denominator (450 reads) and found that the effect persisted (Fig. 4B, right). Indel profiles from combined control clones transfected with the strong gRNA had a higher frequency of larger deletions (5 bp deletions and larger), and correspondingly lower frequency of small indels, than clones transfected with the low efficiency gRNA (Fig. 4C). This shift was reminiscent of one observed in NHEJ-deficient clones (such as Lig4 and Xrcc4) and could be interpreted as a relative increase in DNA resection and MMEJ-activity. The magnitude of the effect was small (no indel size changed in frequency by more than 5% percentage points), but reproducible between biological replicates.

There was considerable variability in the mutagenic efficiency among control clones in the flow cytometric assays (Fig. 1B). We speculated that these differences in control and non-significant clones will correlate with differences between indel profiles. To ensure the highest dynamic range, we used day 14, gRNA #48U flow cytometry samples as a gauge, since in this sample only about 19% of the control cells are mutagenized. We found that mutagenic efficiency in this sample measured by flow cytometry correlated with divergence in indel profiles, as evidenced by the separation of clones in principal component space (Fig. 4D). We concluded higher mutagenic efficiency of Cas9 pushes the DNA repair process towards more mutagenic, MMEJ-like outcomes.

Discussion

We investigated the consequences of Cas9 mutagenesis in a panel of homogenous mouse embryonic stem cells deficient for DNA damage repair genes. We found that the frequency of the complex lesions and large deletions (>260 bp) is increased by NHEJ deficiency and decreased by deficiencies in resection and MMEJ repair (*Nbn* and *Polq*). Large deletion frequency correlated with the increase in extent of microhomology and size of small indels. These result are consistent with the described functions of the identified genes (e.g. ref. [16]) and imply a continuity of underlying repair processes across large genomic distances. They also underscore the potential mutagenic danger of NHEJ inhibition, a common strategy for increasing the frequency of templated repair.

Our results also imply a strategy for decreasing the frequency of complex lesions, namely inhibition of MMEJ or resection, in particular by targeting *Nbn*. While global inhibition will decrease cellular viability and, in case of resection, genome-wide repair fidelity (by preventing homologous recombination), a more targeted approach may be viable, like combining Cas9 enzyme with a resection-inhibiting moiety. Combining this strategy with prime editing could further reduce damage in rare cases when a DSB occurs. Another potential application of resection inhibition is to expedite the production of engineered cell clones by reducing the incidence of cryptic complex lesions[39]. Moreover, repair outcomes in a resection-deficient context are much more predictable and more likely to lead to frame-disrupting 1–2 bp deletions and insertion. However, we note that the frame-shifting phenotype caused by 1 bp lesions may preserve some physicochemical properties of some proteins[50] and thus not always result in the desired null phenotype.

A number of deficiencies in auxiliary repair genes had locus-selective effects. This observation is fully consistent with the well-described substrate-specificity of the repair genes involved, and the fact local target sequence and chromatin state shape repair outcomes. Combining local nuclease-coupled manipulation of DNA repair machinery (e.g. ref. [28]) and indel profile prediction may be a viable strategy for obtaining the desired editing outcomes at a wide range of targets with minimal disruption to physiological DNA repair. Our observations also suggest

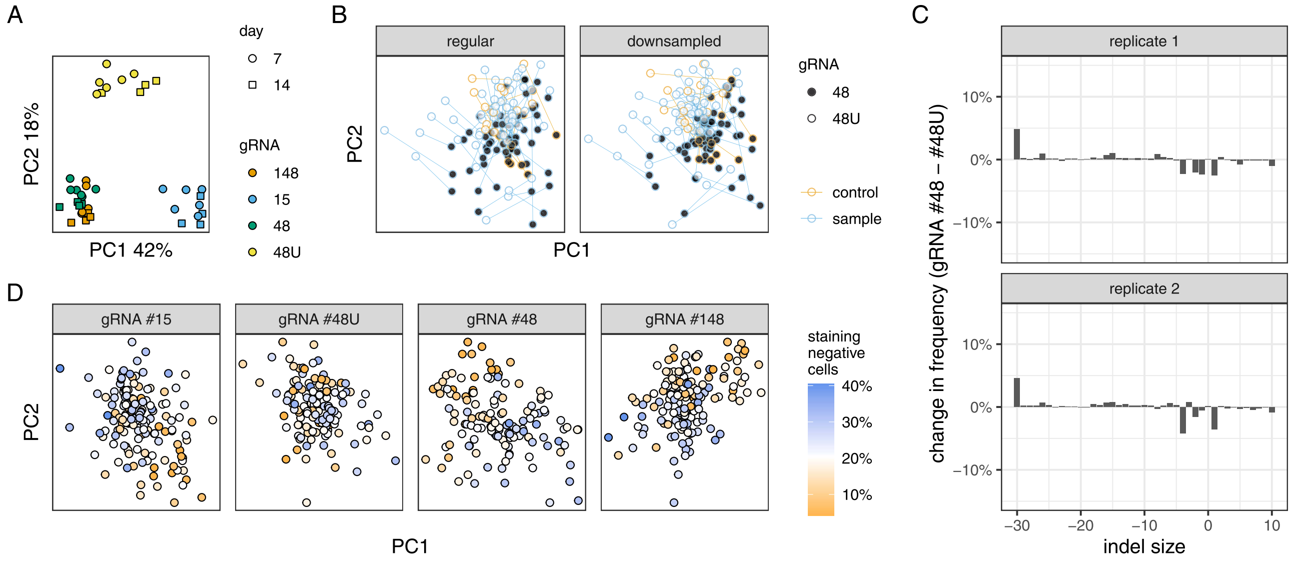

Fig. 4 Efficiency of mutagenesis affects DNA repair outcomes. A Relationship between flow cytometry samples. gRNA #48U is a weaker version of #48. $N = 4$–6 biologically independent replicates. **B** Relationship between clones based on their indel profiles, analogous to Fig. 2B. Only non-significant clones are shown for clarity. Biological replicates ($N = 2$) were averaged. **C** Difference in indel frequency between the regular gRNA #48 and its less active counterpart #48U. Same display conventions as in Fig. 2. **D** Relationship between clones based on their indel profiles, analogous to Fig. 2B. Each clone is colored by the frequency of mutagenesis assayed by flow cytometry on day 14 using gRNA #48U, a proxy for Cas9 activity. For clarity, only non-significant clones are depicted.

additional repair phenotypes may be discovered when the range of targets is expanded.

We were surprised to discover that *Trex1* deficiency had altered indel profiles. Trex1, discovered in 1969 and purified three decades later, has been studied extensively for its role in preventing autoimmunity caused by excess of ssDNA in the cytosol[51–55]. Its in vitro exonuclease activity is fully compatible with a role in DSB repair but, to our knowledge, this activity was hitherto unknown. The underlying mechanism is unknown. Trex1 could potentially act upon DSB directly, for example during S phase, when it is involved in resolution of dicentric chromosomes[56,57]. Alternatively, the observed effect could be a secondary consequence of ssDNA accumulation, perhaps related to the increase in mutagenic repair upon transfection of non-homologous DNA reported by Richardson et al.[58].

We have shown that increased efficiency of Cas9-mediated mutagenesis correlated with MMEJ-like shift towards larger indels. We have noticed a similar effect before, when comparing different modes of Cas9 delivery[59]. Multiple mutually non-exclusive causes for this are possible. More efficient Cas9 complex can recut the DNA after perfect repair sooner (potentially leading to chromatin state dependent repair modulation), cut both sister chromatids simultaneously more often (confounding HR repair), stay bound to DNA after introducing the cut longer on the average (ref. [60], perhaps interfering with the assembly of repair machinery or causing replication fork stalling or collapse) and exert its potential exonuclease activity more intensely (in vitro:[61,62]), than Cas9 of lower efficiency. Finally, the difference in observed profiles could in part be a temporary consequence of slower MMEJ repair dynamics. As long as additional DSBs are being introduced, there is an excess of alleles under repair by the MMEJ pathway compared to the faster NHEJ pathway. Alleles in the process of being repaired cannot be amplified and are thus depleted from the observed indel profile. It is not trivial to figure out in which direction this process would push the indel profile, and what the magnitude of this effect is in our assay. An experiment using inducible gRNAs or inducible Cas9 of different strength would clarify this issue. However, we believe it is unlikely that the difference we observed is entirely due to this, as the effect persist, when mutagenesis is nearing saturation (Supplementary Fig. 2, gRNAs #15, #48 and #148). Our results warrant further investigation and urge caution when using high concentrations of nucleases.

The observation that increased mutagenesis pushes repair towards MMEJ-like outcomes suggest that off-target sites, which are mutagenized in a less efficient fashion, will also exhibit fewer large indels than on-target sites. However, the difference between target sequences would certainly be confounding and therefore any such relation can only be verified to be true on the average by studying multiple on-target/off-target groups.

Despite deriving all our clones simultaneously from a pure, single cell cloned line, we observed a variability in mutagenic efficiency between control clones (e.g. 19–63% on day 7 with gRNA#48). The initial round of subcloning likely removed most of the genetic variability, both genomic and related to individual lentiviral transductions (reverse transcription and APOBEC-mediated mutagenesis). Therefore, we speculate that differences in efficiency were due to a stochastic, mitotically heritable, epigenetic process acting on the Cas9 transgene, possibly position effect variegation. Since varying intensity of DSB introduction has influence on indel profile measurements, this might have precluded us from observing more subtle changes brought about by DNA repair deficiencies. Variation in mutagenic efficiency between Cas9 clones needs to be carefully consider as a potential confounder, when studying DNA damage repair.

Many genes in our library had no clones with statistically significant changes in indel profiles. Since their knock-out is only presumed based on the absence of small, frame-preserving indels (with the exception of *Parp1* and *Lig1*, which were confirmed by Western blot), we cannot claim it as evidence of no function. We note that some NGS-validated clones deficient for core end joining genes, such as Xlf#1, Xrcc4#3 and Polq#1 exhibited a very mild or completely absent phenotype, while other clones with similar genotypes had very strong phenotypes. This, combined with the fact no significant effect was observed for genes with well-described functions in end joining, such as *Parp1*, *Lig1*, *Lig3*, *Ctip* and *Paxx*, implies genetic compensation might play a role.

The quick pace of development of genetic therapies, which may in principle involve any human tissue at any stage of development, makes context-dependent study of DNA repair particularly important. By demonstrating that deficiency in a single repair gene may alter the frequency of large deletions and complex lesions by an order of magnitude, and that requirements for repair machinery can be highly target dependent, we point out the gaps in our knowledge that need to be filled to enable efficient and safe gene therapy.

Methods

Generation of the Cas9+ embryonic stem cell clone. An EF1a-Cas9-T2A-blastR transgene in a pKLV backbone[2,63] was introduced by lentiviral transduction into a highly heterozygous CB9 mouse embryonic stem cell line, derived from a cross between CAST and C57BL/6 strains[64]. Low titre of the virus was used to achieve low copy number. Blasticidin selected, single cell cloned colonies were isolated and tested for Cas9 efficiency using a flow cytometric assay with self-targeting BFP-GFP-anti-GFP construct[2] or a gRNA against *Cd9* gene[31]. The most efficient and homogenous clone "CBA9" was picked for library creation (Supplementary Fig. 4A).

All ES cells used in this study were propagated on SNL-blast feeder cells resistant to neomycin and blasticidin or SNL-HBP feeders resistant to neomycin, blasticidin, hygromycin and puromycin. SNL-HBP were created for this purpose by stable transposition of a low passage SNL cell line with a *PiggyBac* PGK-hygroR-blastR-puroR construct using hyperactive *PiggyBac* transposase[65] and selecting a multi-resistant pool of cells. Mouse ES cells grown on SNL-HBP feeders were found to have no morphological abnormalities compared to those grown on SNL feeder cells (data not shown).

Generation of the DNA damage repair deficient library. *PiggyBac* transposons expressing a hygromycin resistance gene and gRNAs against DNA damage repair genes (Supplementary Data, guides tab; control, knock-out and knock-out-fail gRNAs) were introduced into CBA9 Cas9+ cells in an arrayed format using lipofection. Cells were selected for stable integration using 140 μg/ml hygromycin and single cell cloned. gRNA-targeted loci were amplified using barcoded primers (Supplementary Data, guides tab) and sequenced using MiSeq. Mutagenic alleles were called using CRISPResso2[66] and manual curation of reads aligned using STAR[67]. The latter method yielded additional large deletion alleles (>50 bp) missed by CRISPResso2.

Based on the recovered genotype, clones were classified as "perfect", "in-frame" or "good". Clones whose all detectable alleles were frame-disrupting and, where applicable, could be assigned to a strain (one BL6 and one CAST allele), were deemed "perfect". These are very likely to have lost target gene function. Clones containing any alleles likely to be functional (frame-preserving insertion or deletions smaller than 30 bp) were considered "in-frame", unlikely to have lost gene function, unless a critical protein domain was affected. Other clones, including those with more than two alleles, with one allele at loci without strain specific SNPs (potentially homozygous, or harboring a complex lesion undetectable by short-range PCR), with any in-frame deletions 30 bp or larger (likely deleterious) or with alleles that could not be assigned to a strain at a heterozygous locus (because the lesion erased the distinguishing SNPs), were classified as "good". Control clones were obtained using various non-targeting gRNAs and a gRNA targeting a safe harbor *Rosa26* locus. In total, 57 perfect, 18 good and 8 in-frame experimental clones, along with 12 controls, were included in the final library for a total of 95 clones (Supplementary Fig. 4B). One well was intentionally left empty as a negative control for cell and DNA carry-over. The library contained NGS-validated clones deficient for 32 genes, with 1–4 clones for each gene (Supplementary Data, clones tab). Two of these genes, *Brca2* and *Xrcc1*, were only represented by "in-frame" clones. No clones were included for seven other targeted genes, which yielded no promising candidates (*Brca1*, *Exo1*, *Mre11a*, *Rif1*, *Rnaseh2a*, *Fen1* and *Mad2l2*).

Western blots were performed using following antibodies: rabbit Parp1 (ab191217, abcam, dilution 1:8000), rabbit Lig1 (18051-1-AP, Proteintech, 1:500), rabbit Nbs1 (A301-284A, Bethyl, 1:2000), rabbit Xlf (A300-730A, Bethyl, 1:2000), mouse Actin (SC-47778, Santa Cruz, 1:200), HRP goat anti-rabbit antibody (ab205718, abcam, 1:2000) and HRP goat anti-mouse (ab205719, abcam, 1:4000) following manufacturers' recommendations.

Flow cytometric assessment of mutagenesis efficiency and complex lesion frequency. Flow cytometric assays were conducted as previously described[31,44]. In short, *PiggyBac* transposons expressing one of the five experimental gRNAs (#15, #33, #48, #48U and #148) and a puromycin resistance gene were introduced into the library clones in an arrayed format using lipofection. Cells were selected for stable integration using 10 ng/μ l puromycin. This strategy ensures a near complete mutagenesis. On day 7 and day 14 post-transfection, cells were stained using FLAER reagent (for PigA activity; Cedarlane) or Itga6-PE antibodies (#313612, Biolegend) and analyzed using Cytoflex flow cytometer and its native software (Beckman-Coulter). Six replicates were performed for day 7 and four for day 14. Gating strategy is in Supplementary Fig. 4C.

Data was extracted, processed and visualized in R, using packages flowCore, flowWorkspace, openCyto and ggcyto[68–70]. The same gating was used throughout, except in replicate 6 on day 7, in which cells had to be restained and gates adjusted to lower staining efficiency. Size gating removed feeder cells, which are much larger than mouse ES cells, as confirmed by very low number of events in the empty control well (data not shown). A bacterial infection was detected in replicate 1 on day 14 - cells were processed as usual and data was retained. Raw percentages of staining-positive cells from each plate (that is, replicate, experimental gRNA, staining combination) were mean and standard deviation normalized and resulting raw z-scores were decomposed using PCA. Principal components numbers one and two captured 60% of variation and separated the samples by gRNA and time of sampling. The next two components separated two batches of replicates (Supplementary Fig. 5A). These batches were initiated from different master plates and used different lots of some reagents. The second batch grew faster (data not shown) and, possibly as a consequence, experienced an overall increased level of mutagenesis (Supplementary Fig. 5B). We only used data derived from principal components numbers one and two for analysis, expressed as a z-score with relation to mean and standard deviation of the control samples (e.g. in Fig. 1B). Raw frequencies of gene expression negative cells, raw z-scores and PCA-regressed z-scores are presented side-by-side in Supplementary Fig. 2.

Analysis of indel profiles. For the purpose of indel profile analysis, cells were passaged at least twice without feeder cells on gelatin in medium supplied with LIF, prior to DNA extraction. Loci targeted with five experimental gRNAs in 95 clones in two biological replicates (day 14, replicates 5 and 6 in the flow cytometric assays) were amplified using barcoded primers (Supplementary Data, guides tab; amplicons of 244–283 bp) and sequenced using MiSeq. Demultiplexed reads were handled as described in Allen et al.[5]. In brief, reads were transformed into indel signatures characterised by their size, type and location with respect to the cut site (3rd/4th nucleotide 5' of the PAM). For example, "D10_L-13C2R0" is a 10 bp deletion ("D"), whose last unmodified nucleotides are thirteen to the left of the cut site and at the cut site. Two nucleotides of microhomology ("M") could map at either end of that interval. When an indel contains insertions ("I"), microhomology indicates that this part of the insertion matches at either end of the interval, indicating a possible templated insertion.

Samples transfected using control gRNA #33 targeting GFP were found to contain negligible amounts of indels at #15, #48 and #148 sites and were discarded (data not shown). The following filters were applied to the remaining 760 samples. Indels that would result in loss of more than 150 bp were filtered out to avoid primer-dimers (0.6% read loss) and samples with fewer than 400 remaining reads were removed (18/760 samples lost; each gRNA-clone combination retained at least one sample). Symmetrized Kullback-Leibler divergence ("KL") was calculated for all pairs of samples as described in Allen et al.[5]. The resulting KL matrix was decomposed using multidimensional-scaling (MDS) for visualization and statistical testing. A bivariate normal distribution was fitted to controls using principal components numbers one and two of the MDS-decomposed KL matrix (Supplementary Fig. 3A) and the associated *p* value for each sample was derived from chi-squared distribution (since for the bivariate standard normal, the squared distance of a random point from the mean has a chi-squared distribution with two degreees of freedom). An indel profile of a clone was considered significantly different from controls, if all of its replicates had a FDR-corrected $p < 0.01$. Analysis and visualization were performed in R using ggplot2, ggrepel (https://ggrepel.slowkow.com/) and tidyverse group of packages, as well as colorblindr and cowplot[71,72].

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

Data availability

All data necessary for recreating the figures are available at https://gitlab.com/lotard/medraka_paper. Any additional data are available on request.

Code availability

All scripts necessary for recreating the figures are available at https://gitlab.com/lotard/medraka_paper.

Received: 5 August 2020; Accepted: 28 April 2022;
Published online: 14 June 2022

References

1. Sfeir, A. & Symington, L. S. Microhomology-mediated end joining: a back-up survival mechanism or dedicated pathway? *Trends biochemical Sci.* **40**, 701–714 (2015).

2. Koike-Yusa, H., Li, Y., Tan, E.-P., Del Castillo Velasco-Herrera, M. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* **32**, 267–73 (2014).
3. van Overbeek, M. et al. DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks. *Mol. Cell* **63**, 633–646 (2016).
4. PienTan, E., Li, Y., Martin, Velasco-Herrera, 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).
5. Allen, F. et al. Predicting the mutations generated by repair of Cas9-induced double-strand breaks, *Nat. Biotechnol.*, nov 2018. ISSN 1087-0156. https://doi.org/10.1038/nbt.4317 (2018).
6. Shen, M. W. et al. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* **563**, 646–651 (2018).
7. Chen, W. et al. Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair, *Nucl. Acids Res.*, jun. ISSN 0305-1048. https://doi.org/10.1093/nar/gkz487 (2019).
8. Leenay, R. T. et al. Large dataset enables prediction of repair after CRISPR-Cas9 editing in primary T cells. *Nat. Biotechnol.* **37**, 1034–1037 (2019).
9. Lemos, B. R. et al. CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles, *Proc. Natl Acad. Sci.*, 201716855. ISSN 0027-8424. https://doi.org/10.1073/pnas.1716855115 (2018).
10. Shi, X. et al. Cas9 has no exonuclease activity resulting in staggered cleavage with overhangs and predictable di- and tri-nucleotide CRISPR insertions without template donor. *Cell Discov.* **5**, 53 (2019).
11. Boulton, S. J. & Jackson, S. P. Saccharomyces cerevisiae Ku70 potentiates illegitimate DNA double-strand break repair and serves as a barrier to error-prone DNA repair pathways. *EMBO J.* **15**, 5093–103 (1996).
12. Liang, Z., Sunder, S., Nallasivam, S. & Wilson, T. E. Overhang polarity of chromosomal double-strand breaks impacts kinetics and fidelity of yeast non-homologous end joining. *Nucleic Acids Res.* **44**, 2769–2781 (2016).
13. Wang, M. et al. PARP-1 and Ku compete for repair of DNA double strand breaks by distinct NHEJ pathways. *Nucleic Acids Res.* **34**, 6170–82 (2006).
14. Yu, A. M. & McVey, M. Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. *Nucleic Acids Res.* **38**, 5706–17 (2010).
15. Schimmel, J., Kool, H., van Schendel, R. & Tijsterman, M. Mutational signatures of non-homologous and polymerase theta-mediated end-joining in embryonic stem cells, *EMBO J.*, page e201796948. ISSN 0261-4189. https://doi.org/10.15252/embj.201796948 (2017).
16. Carvajal-Garcia, J. et al. Mechanistic basis for microhomology identification and genome scarring by polymerase theta. *Proc. Natl Acad. Sci.* **117**, 8476–8485 (2020).
17. Jasin, M. & Rothstein, R. Repair of strand breaks by homologous recombination. *Cold Spring Harb. Perspect. Biol.* **5 VN - re11**, 1–19 (2013).
18. Roth, T. L. et al. Reprogramming human T cell function and specificity with non-viral genome targeting, *Nature*, page 1, jul 2018. ISSN 0028-0836. https://doi.org/10.1038/s41586-018-0326-5 (2018).
19. Martin, R. et al. Selection-free, high frequency genome editing by homologous recombination of human pluripotent stem cells using Cas9 RNP and AAV6, *bioRxiv*, 252163, jan. https://doi.org/10.1101/252163 (2018).
20. Maruyama, T. et al. Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nat. Biotechnol.* **33**, 538–542 (2015).
21. Chu, V. T. et al. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat. Biotechnol.* **33**, 543–8 (2015).
22. Riesenberg, S. & Maricic, T. Targeting repair pathways with small molecules increases precise genome editing in pluripotent stem cells. *Nat. Commun.* **9**, 2164 (2018).
23. Riesenberg, S. et al. Simultaneous precise editing of multiple genes in human cells, *Nucl. Acids Res.*, aug 2019. ISSN 0305-1048. https://doi.org/10.1093/nar/gkz669 (2019).
24. Vertex ramps up CRISPR repair, Mar 2019. ISSN 1546-1696. [Online; accessed 10. May 2020].
25. Nambiar, T. S. et al. Stimulation of CRISPR-mediated homology-directed repair by an engineered RAD18 variant. *Nat. Commun.* **10**, 3395 (2019).
26. Paulsen, B. S. et al. Ectopic expression of RAD52 and dn53BP1 improves homology-directed repair during CRISPR-Cas9 genome editing. *Nat. Biomed. Eng.* **1**, 878–888 (2017).
27. Song, J. et al. RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. *Nat. Commun.* **7**, 1–7 (2016).
28. Rees, H. A., Rees, H. A., Yeh, Wei-Hsi & Liu, D. R. Development of hRad51-Cas9 nickase fusions that mediate HDR without double-stranded breaks. *Nat. Commun.* **10**, 2212 (2019).
29. Wienert, B. et al. Timed inhibition of CDC7 increases CRISPR-Cas9 mediated templated repair. *Nat. Commun.* **11**, 2109 (2020).
30. Shin, H. Y. et al. CRISPR/Cas9 targeting events cause complex deletions and insertions at 17 sites in the mouse genome. *Nat. Commun.* **8**, 1–10 (2017).
31. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.* **36**, 765 (2018).
32. Adikusuma, F. et al. Large deletions induced by Cas9 cleavage. *Nature* **560**, E8–E9 (2018).
33. Nelson, C. E. et al. Long-term evaluation of AAV-CRISPR genome editing for Duchenne muscular dystrophy. *Nat. Med.* **25**, 427–432 (2019).
34. Mianné, J. et al. Correction of the auditory phenotype in C57BL/6N mice via CRISPR/Cas9-mediated homology directed repair. *Genome Med.* **8**, 16 (2016).
35. Renaud, J. B. et al. Improved genome editing efficiency and flexibility using modified oligonucleotides with TALEN and CRISPR-Cas9 nucleases. *Cell Rep.* **14**, 2263–2272 (2016).
36. Birling, M. C. et al. Efficient and rapid generation of large genomic variants in rats and mice using CRISMERE. *Sci. Rep.* **7**, 43331 (2017).
37. Codner, G. F. et al. Application of long single-stranded DNA donors in genome editing: generation and validation of mouse mutants. *BMC Biol.* **16**, 70 (2018).
38. Cullot, G. et al. CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations. *Nat. Commun.* **10**, 1136 (2019).
39. Weisheit, I. et al. Detection of deleterious on-target effects after HDR-mediated CRISPR editing, *bioRxiv*, page 2020.03.27.012104, mar 2020. https://doi.org/10.1101/2020.03.27.012104 (2020).
40. Ikeda, K. et al. Efficient scarless genome editing in human pluripotent stem cells. *Nat. Methods* **15**, 1045–1047 (2018).
41. Rayner, E. et al. CRISPR-Cas9 Causes Chromosomal Instability and Rearrangements in. *Cancer Cell Lines, Detectable Cytogenetic Methods, CRISPR J.* **2**, 406–416 (2019).
42. 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* **61**, 5985–91 (2016).
43. Anzalone, A.V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA, *Nature*, pages 1–1, oct 2019. ISSN 0028-0836. https://doi.org/10.1038/s41586-019-1711-4 (2019).
44. Kosicki, M. Cas9-induced on-target genomic damage, 2019.
45. Le Chalony, C. et al. Partial complementation of a DNA ligase I deficiency by DNA ligase III and its impact on cell survival and telomere stability in mammalian cells. *Cell. Mol. Life Sci.* **69**, 2933–2949 (2012).
46. Shou, J., Li, J., Liu, Y. & Wu, Q. Precise and predictable CRISPR Chromosomal Rearrangements Reveal Principles of Cas9-Mediated Nucleotide Insertion. *Mol. Cell* **71**, 498–509.e4 (2018).
47. Roidos, P. et al. A scalable CRISPR/Cas9-based fluorescent reporter assay to study DNA double-strand break repair choice. *Nat. Commun.* **11**, 4077 (2020).
48. Kent, T., Chandramouly, G., McDevitt, S. M., Ozdemir, A. Y. & Pomerantz, R. T. Mechanism of microhomology-mediated end-joining promoted by human DNA polymerase *θ*. *Nat. Struct. Mol. Biol.* **22**, 230–7 (2015).
49. Chang, H. H. Y., Pannunzio, N. R., Adachi, N. & Lieber, M. R. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat. Rev. Mol. Cell Biol.* **18**, 495–506 (2017).
50. Bartonek, L., Braun, D. & Zagrovic, B. Frameshifting preserves key physicochemical properties of proteins. *Proc. Natl Acad. Sci.* **117**, 5907–5912 (2020).
51. Lindahl, T., Gally, J. A. & Edelman, G. M. Properties of deoxyribonuclease 3 from mammalian tissues. *J. Biol. Chem.* **244**, 5014–9 (1969).
52. Höss, M. et al. A human DNA editing enzyme homologous to the Escherichia coli DnaQ/MutD protein, *EMBO J.*, 1999. ISSN 02614189. https://doi.org/10.1093/emboj/18.13.3868 (1999).
53. Mazur, D. J. & Perrino, F. W. Identification and expression of the TREX1 and TREX2 cDNA sequences encoding mammalian 3'-5' exonucleases, *J. Biol. Chem.* **274**, 19655–19660 (1999).
54. Crow, Y. J. et al. Mutations in the gene encoding the 3'-5' DNA exonuclease TREX1 cause Aicardi-Goutières syndrome at the AGS1 locus. *Nat. Genet.* **38**, 917–920 (2006).
55. Stetson, D. B., Ko, J. S., Heidmann, T. & Medzhitov, R. Trex1 Prevents Cell-Intrinsic Initiation of Autoimmunity. *Cell* **134**, 587–598 (2008).
56. Maciejowski, J., Chatzipli, A., Dananberg, A., de Lange, T., & Campbell, P. APOBEC3B-dependent kataegis and TREX1-driven chromothripsis in telomere crisis, *bioRxiv*, 725366, 2019. https://doi.org/10.1101/725366 (2019).
57. Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J. & De Lange, T. Chromothripsis and Kataegis Induced by Telomere Crisis. *Cell* **163**, 1641–1654 (2015).
58. Richardson, C. D., Ray, G. J., Bray, N. L. & Corn, J. E. Non-homologous DNA increases gene disruption efficiency by altering DNA repair outcomes. *Nat. Commun.* **7**, 1–7 (2016).
59. Kosicki, M. et al. Dynamics of indel profiles induced by various CRISPR/Cas9 delivery methods. *Prog. Mol. Biol. Transl. Sci.* **152**, 49–67 (2017).

60. Sternberg, S. H., SyRedding, Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
61. Jinek, M. et al. A Programmable Dual-RNA - Guided DNA Endonuclease in Adaptice Bacterial Immunity, *Science* (New York, N.Y.) **337**, 816–822 (2012).
62. Stephenson, A. A., Raper, A. T. & Suo, Z. Bidirectional Degradation of DNA Cleavage Products Catalyzed by CRISPR/Cas9. *J. Am. Chem. Soc.* **140**, 3743–3750 (2018).
63. Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
64. Strogantsev, R. et al. Allele-specific binding of ZFP57 in the epigenetic regulation of imprinted and non-imprinted monoallelic expression. *Genome Biol.* 1–18 (2015).
65. Yusa, K. et al. Targeted gene correction of α 1-antitrypsin deficiency in induced pluripotent stem cells. *Nature* **478**, 391–4 (2011).
66. Clement, K. et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis, mar 2019. ISSN 15461696. (2019).
67. Dobin, A. et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
68. Van, P., Jiang, W., Gottardo, R. & Finak, G. ggCyto: next generation open-source visualization software for cytometry. *Bioinforma. (Oxf., Engl.)* **34**, 3951–3953 (2018).
69. Finak, G. et al. OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Computational Biol.* **10**, e1003806 (2014).
70. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/, 2013. ISSN 16000706. (2013).
71. Wickham, H. et al. Welcome to the Tidyverse, *J. Open Source Software*. **4**, 1686 (2019).
72. Claus O. Wilke. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O'Reilly Media, 671 Apr 2019. ISBN 978-149203108-6. (2019).

Acknowledgements

We would like to thank Lara Urban for help with running the PEER package, which in our case yielded the same results as PCA. Ilias Georgakopoulos-Soares, Ozdemirhan Serçin and Michał Barski offered useful comments on the manuscript. The bioRxiv version of the manuscript was typeset using a slightly modified Henriques Lab LaTeX template. This work is funded by Wellcome Trust grant no. 098051.

Author contributions

M.K. designed, executed, analyzed and interpreted the experiments and wrote the paper. A.B. supervised the project and contributed to writing of the manuscript. F.A. processed indel profile data and made useful comments on the manuscript. K.T. supervised Western blot experiments, while F.S. and Y.P. executed them.

Competing interests

The authors declare no competing interests. F.A. is currently employed at Genomics PLC, UK, but her involvement in the work reported in this paper was all undertaken at the Wellcome Sanger Institute.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-30480-8.

Correspondence and requests for materials should be addressed to Allan Bradley.

Peer review information *Nature Communications* thanks Nariman Battulin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.