Predicting the mutations generated by repair of Cas9-induced double-strand breaks

Felicity Allen[#@,1], **Luca Crepaldi**[#1], **Clara Alsinet**[1], **Alexander J. Strong**[1], **Vitalii Kleshchevnikov**[1], **Pietro De Angeli**[1], **Petra Palenikova**[1], **Anton Khodak**[1], **Vladimir Kiselev**[1], **Michael Kosicki**[1], **Andrew R. Bassett**[1], **Heather Harding**[2], **Yaron Galanty**[3,4], **Francisco Muñoz-Martínez**[3,4], **Emmanouil Metzakopian**[1,5], **Stephen P. Jackson**[3,4], **Leopold Parts**[@,1,6]

[1]Wellcome Sanger Institute, Hinxton, United Kingdom

[2]Cambridge Institute of Medical Research, University of Cambridge, Cambridge, United Kingdom

[3]The Wellcome/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, United Kingdom

[4]Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

[5]UK Dementia Research Institute, Cambridge, United Kingdom

[6]Department of Computer Science, University of Tartu, Tartu, Estonia

[#] These authors contributed equally to this work.

Abstract

The DNA mutation produced by cellular repair of a CRISPR/Cas9-generated double-strand break determines its phenotypic effect. It is known that the mutational outcomes are not random, and depend on DNA sequence at the targeted location. Here we systematically study the influence of flanking DNA sequence on repair outcome by measuring the edits generated by >40,000 guide RNAs in synthetic constructs. We performed the experiments in a range of genetic backgrounds and using alternative CRISPR/Cas9 reagents. In total, we gathered data for >10^9 mutational outcomes. The majority of reproducible mutations are insertions of a single base, short deletions, or longer microhomology-mediated deletions. Each gRNA has an individual cell-line dependent bias toward particular outcomes. We uncover sequence determinants of the produced mutations,

[@]Correspondence to LP (leopold.parts@sanger.ac.uk) or FA (fa9@sanger.ac.uk).

Data availability

Raw sequence data is available at European Nucleotide Archive (Project accession PRJEB12405 / ERP013879; sample accessions provided in Dataset 2). Full code is available at https://github.com/felicityallen/SelfTarget, as a runnable Code Ocean module (https://doi.org/10.24433/CO.6bc7bcae-d736-475b-bae5-00ca0562d401) and as Manuscript Related File 1. Processed mutational profiles are provided on https://figshare.com/articles/processed_mutational_profiles/7312067 (https://doi.org/10.6084/m9.figshare.7312067)

Author Contributions

FA: designed experiments, analysed data, wrote paper. LC: designed experiments, performed experiments, wrote paper. CA: performed experiments in human iPSCs. AJS, EM: performed experiments in mouse ESCs. VKle: analysed data, wrote paper. AK, VKi: created web server. PDA, PP: performed experiments. MK, ARB: generated TREX2 constructs. HH: generated CHO-Cas9 line. YG, FM-M, SPJ: generated RPE-1-Cas9 and HAP1-Cas9 lines. LP: designed experiments, contributed to data analysis, wrote paper. All authors contributed to drafting the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts

and use these to derive a predictor of Cas9 editing outcomes. Improved understanding of sequence repair will allow better design of gene editing experiments.

Introduction

CRISPR/Cas9 is a transformative DNA editing technology[1]. It operates by recruiting the Cas9 nuclease to a genomic locus with a protospacer-adjacent motif (PAM) using a short synthetic guide RNA (gRNA) with a 18-20 nt sequence matching the desired target. Cas9 then cuts DNA at that location, and when the double strand break is repaired by cellular machinery, frameshift mutations can occur, disabling translation of the correct protein.

Cas9-generated mutations result from imperfect action of DNA repair pathways that are activated to remedy the double strand break. The major repair mechanisms include non-homologous end joining, which re-ligates the generated ends, often introducing errors of a few nucleotides; and microhomology-mediated end joining, in which short tracts of local matching sequence anneal, ultimately resulting in deletion of the intervening bases[2,3]. Choice of pathway is influenced by a host of factors including cell cycle stage, and availability of repair enzymes[4,5]. It has been shown, that the frequency of alternative Cas9 editing outcomes ("the mutational profile"; Figure 1) is largely reproducible, and depends on the targeted sequence[6–10], indicating that the errors in repair occur in a non-random manner. Although DNA repair pathways and their key components have been characterized, the biases that favor one mutation over another are not fully understood, especially for the breaks inflicted by Cas9.

To date, mutational profiles have not been measured at scale. The main barrier has been the labor necessary to individually amplify the sequence at each of the targeted loci. The largest current dataset of genomic repair profiles comprises 436 profiles examining 96 unique gRNA sequences using the Cas9 protein from *S. pyogenes*[7], recently followed up with studies of additional target sites[11,12]. More gRNAs (~1,400) were employed in a study that introduced the target and gRNA into cells simultaneously[13], but the low probability of a gRNA and its corresponding target meeting in the same cell resulted in an average mutation rate of 0.2%, and insufficient data for a comprehensive analysis. An approach introducing gRNA and target in the same synthetic construct has been used for the Cpf1 nuclease[14], and the *Staphylococcus aureus* Cas9 enzyme[15]. Both profiled proteins have a shorter RNA scaffold sequence, enabling a simpler library cloning procedure to assess more gRNAs. However, differences between both the proteins themselves and the characteristics of the DNA breaks they generate mean that these results are not directly applicable to Cas9. Whereas the Cpf1 data was used to develop an algorithm that predicts indel frequencies, no attempt was made to predict the SaCas9-generated mutation frequencies or the exact repair outcomes for either of the enzymes.

Here, we present a large-scale measurement of Cas9-generated gRNA repair profiles. We synthesized over 40,000 DNA constructs, each containing both a gRNA and its target, introduced them into Cas9-expressing cell lines, and sequenced the targeted loci. We confirm that our measurements are informative of events at endogenous sites, describe the

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

dominant outcomes and their sequence dependence in a range of cell lines, and present an accurate predictive model for forecasting the outcomes of an edit.

Results

Measuring repair outcomes *en masse*

The main hurdle in measuring a large number of repair outcomes is the need to selectively amplify each targeted locus. To circumvent this, we designed a construct that encodes a gRNA expression cassette together with its 23nt PAM-endowed target sequence within a larger 79nt variable context, and flanked by common PCR priming sites on both sides (Figure 1, S1). The variable context allowed us to systematically change the local sequence to directly test its influence on the repair outcome, and to unambiguously assign each sequenced target to its gRNA-target pair of origin. Using high throughput oligonucleotide synthesis followed by custom cloning reactions (Online Methods), we generated several libraries of gRNA-target pairs, with a total of 41,630 constructs. We delivered these into cells using lentiviral infection at 0.5-0.6 multiplicity (Table S1), cultured for seven days to ensure saturated editing while avoiding drift (Figure S2-S4), then isolated genomic DNA, amplified the target sequence in its context, and sequenced at high coverage (Figure S5) to measure the frequency of insertions and deletions that had occurred.

Synthetic repair profiles are reproducible, and faithfully capture endogenous repair outcomes

First, we demonstrate that our measurements are sequence-specific and reproducible in the human chronic myelogenous leukemia (K562) cell line. Here and elsewhere, we use the symmetric Kullback-Leibler ("KL") divergence, a natural information-theoretic metric related to relative entropy of probability distributions, to quantify similarity of outcome frequencies (Online Methods). Given adequate read coverage, profiles from biological replicates measuring the same gRNA target were similar, whereas targets of randomly selected gRNAs had markedly different repair outcomes (median KL=0.70 vs 4.8; Figure 2A-C; 6,218 gRNAs from conventional set, Online Methods). The fraction of frameshift edits, a factor that is arguably most important for knock-out experiments, is also highly correlated between biological replicates (Pearson's R=0.9; Figure 2D). Together, these results show that the mutational profiles are reproducible and sequence-specific. Given the negligible influence of whether the conventional[16] or improved[17] gRNA scaffolds were used (median KL=0.77; Figure 2A, 2C), the improved version is employed in all following experiments, unless noted otherwise.

We next tested whether the measurements from our synthetic targets are a good proxy for repair outcomes at endogenous loci. To do this, we took advantage of data from the largest scale study of editing outcomes to date[7], in which 223 human genomic targets for 96 unique gRNAs were individually amplified and sequenced ("endogenous outcomes"). Those 77 of these gRNAs that we were able to successfully clone were included in our library with their genomic contexts (Online Methods). Concordance between synthetic and endogenous outcomes was very good for individual cases (Figure 2A), on the whole (median KL=1.1; Figure 2B), and for recapitulating frameshift edit fraction (Pearson's R=0.78, Figure 2D).

Europe PMC Funders Author Manuscripts

Nevertheless, the observed differences were larger than for biological replicates of our assay, so we inspected the reasons for this. We identified two causes. First, sampling noise due to low sequencing coverage leads to increased divergence (Figure 2E). Second, deletions and rearrangements larger than our measurement size limit of 30nt (Online Methods), and which can be prevalent[18], explain three of the four cases with sufficient read counts that markedly differ (KL > 3). The remaining case ("Overbeek 25") diverges despite high reproducibility between synthetic measurements (Figure S6). Given that our construct only contains 79nt of local context due to limitations of oligonucleotide synthesis, yet produces very similar outcomes for 94% of measured cases with sufficient reads (67/71), this result confirms that sequence surrounding the cut site is the major determinant of Cas9-induced mutational outcomes. We also tested for the influence of chromatin state on the profile similarity, but found that the average divergence between endogenous and synthetic measurements did not differ for endogenous targets in active or repressed chromatin (Table S2).

Repair outcomes in K562 cells are diverse

After concluding that our assay faithfully and reproducibly captures a majority of endogenous mutational outcomes, we surveyed a collection of 6,568 gRNAs that target human genes ("Genomic gRNA-Targets", Online Methods) that we expect to be representative of gRNAs in practical use, in triplicate. We observed that single nucleotide insertions and deletions were most common, with larger insertions occurring only rarely, and shorter deletions favoured over longer ones (Figure 3A). However, a long tail of larger deletion events was present.

Despite shorter deletions being more frequent, most of the Cas9-generated mutations (58%) resulted in a deletion of at least three base pairs (Figure 3B). About half of these (31% of total) occurred between repeating sequences at least 2nt ("microhomology"). Deletions of one or two base pairs made up 18% of all observations, and while insertions of a single base were the most common type of outcome overall (13%), larger insertions were rare (3% of all mutations). More complex outcomes with both insertion and deletion events were present in 8% of measured reads.

Given a similar basal activity of the different DNA repair pathways in all cells of the assay, it is natural to hypothesize that repair outcomes of individual gRNA targets largely conform to the average trend observed above. In fact, there is substantial variability in the relative frequency of different outcome types (Figure 3C). Insertions, single and double nucleotide deletions, and microhomology-mediated deletions can all be present at frequencies ranging from near 0 to over 50% depending on the target, further highlighting the sequence-specific nature of the repair process.

Repair outcomes are biased towards particular alleles. The same specific mutation was most frequent in all three biological replicates for over 60% of gRNAs (Figure 3D), but mutations from different classes were not favored equally. When the consensus existed, it was almost always a single nucleotide insertion (36%), microhomology-mediated deletion of at least 3nt (34%), or a deletion of one or two bases (30%), and could make up over half of all mutations for that gRNA with reproducible frequency (Pearson's R > 0.83; Figure 3C, S7). In contrast, whereas deletions of at least 3nt without microhomology, larger insertions (I>1), and more

Europe PMC Funders Author Manuscripts

complex mutations (I+D) are collectively common (38%, Figure 3B), their frequency for each gRNA is lower and less reproducible (Pearson's R < 0.27; Figure 3C, S7), so that they form less than 1% of the reproducibly most frequent outcomes (Figure 3D).

Overall, half of the measured gRNAs have a single outcome that contributed at least 20% of the observations, and 11% have an outcome that contributed at least 40% (Figure 3E). Yet although on average, the six most frequent alleles per construct account for the majority of its observed mutations, 25 alleles collectively explain only 72% of the data (Figure 3F), indicating a large number of low frequency events. Some of these may be artifactual, but we expect them to form a minority, as we observed an order of magnitude fewer unique mutations in a control experiment lacking Cas9 (Figure S8). Additionally, frequencies of alleles assayed at different time points (seven and 10 days post infection) from the same replicate are more concordant than those of biological replicates (Figure S7), indicating re-sampling of existing low frequency alleles, rather than stochastic measurement noise. Together with evidence of profile reproducibility above, this paints a picture of a complex, yet not completely random repair process for Cas9-generated breaks.

Repair outcomes depend on local sequence properties

Given the reproducible and sequence specific nature of repair outcomes, we next investigated their sequence determinants. For all analyses in this section, we used a larger explorative set of 27,906 constructs that lack a counterpart in the human genome, and cover a broad range of sequence characteristics (Online Methods). Unless noted otherwise in text or figures, all constructs are included in the analysis.

Given that microhomology is known to bias repair of Cas9-induced double strand breaks[6,19], we first systematically evaluated repair outcomes of targets with different microhomology spans (3-15nt), and separating distances (0-20nt). We observed that the fraction of mutations that could be attributed to microhomology-mediated end joining was higher when the matching sequences were separated by shorter distances (Figure 4A). This trend held for all spans of microhomology, but was more pronounced for longer tracts (e.g. Pearson's R=-0.7 for 10 matching bases *vs.* R=-0.2 for three matching bases; Figure 4B).

We next assessed whether imperfectly matching microhomologous sequences also generate corresponding deletions using the gRNAs designed with zero, one or two mismatches in the microhomology region ("Microhomology Mismatch gRNA-Targets"; 571 constructs with mismatches). Indeed, the same alleles are generated at a 30% reduced rate if one mismatch is present, and 50% reduced rate if two (Figure 4C). The presence of mutations in one or the other side of the cut allowed us to further test whether sequence from one side is preferentially retained, but we found no bias either for microhomology-mediated deletions (Figure S9) or the rest (Figure S10). We also observed that sequences with low GC fraction were not as frequently used as repair template as those with higher GC (Figure S11), suggesting preference for a higher melting temperature in the resulting duplex.

There is evidence that single base insertions favor a repeat of the PAM-distal nucleotide adjacent to the cut site in yeast, and to an unknown extent, in humans[9]. In the following, we consider the alleles that are most frequent for a gRNA in all three replicates ("dominant

Europe PMC Funders Author Manuscripts

mutations", Figure 3D). Contrary to the lack of directional bias in deletions (Figure S9, S10), we observed that for over 99% of the 6,572 dominant single base insertions, the PAM-distal nucleotide was repeated (Figure 4D). Further, for 49% of all gRNAs with a thymine as the PAM-distal base at the cut site, insertion of another thymine is the most frequent outcome, whereas this is the case for only 1.6%, 15% and 28% of gRNAs with a guanine, cytosine and adenosine, respectively (Figure 4E).

A similar strong bias is present for small deletions. We observed that 77% of dominant single base deletions correspond to removal of a repeating nucleotide at the cut site (Figure 4F). Deleting one cytosine from a pair was most common (36%), dominating for 30% of 1,843 gRNAs with a cytosine on both sides of the cut. When repeat of another nucleotide was present, its contraction dominated for 12-16% of gRNAs, whereas only 1.6% of gRNAs without a repeat produced a dominant single base deletion (Figure 4G).

Repeat removal was also favored for two-base deletions. Half of dominant dinucleotide deletions contract a repeat of two bases at the cut site, and a further 17% remove a single repeated nucleotide separated by another nucleotide (Figure 4H). In the remaining cases (31%), both bases are removed from one or other side of the cut, but a single base is never removed from both sides. Notably, if a dinucleotide repeat is present at the cut site of a gRNA, its contraction is very likely to be the dominant repair outcome for that gRNA (up to 77% of gRNAs with an AGAG motif; Figure 4I), and preference for the PAM-distal pattern is the opposite of a single base deletion, with thymines giving rise to lower rates (10%, 36/344), while guanines are preferred (62%, 233/377). If alternative sequence configurations are present at the cut site, sequence biases are present (Figure S12), but deleting two bases never dominates for more than 5% of the gRNAs in these other cases (Figure S13).

Mutational outcomes vary with cell line and some Cas9 modifications

Cells can differ in activity of repair processes and/or DNA sequence, both of which influence double strand break repair outcomes[20]. We next performed our assay in human induced pluripotent stem cells (iPSCs), mouse embryonic stem cells (mESCs), Chinese hamster ovary (CHO) epithelial cells, as well as human retina epithelial immortalized cells (RPE-1) and leukemic near-haploid (HAP1) cell lines. We used the same genomic gRNAs as for initial characterization ("Genomic gRNA-Targets" set, Online Methods, e.g. Figure 3), but restricted to the 3,777 with at least 20 mutated reads in all cell lines (median read numbers in Figure 5A).

The overall distribution of repair outcome types was not the same across the different cell lines and organisms studied (Figure 5A), but repair profiles of individual gRNAs remained similar to each other (median KL < 2; Figure 5B, example in Figure S14). Some relative changes of preferred mutation classes were notable. Large insertions occurred more frequently in stem cells, with a 2.6x and 1.7x increase in human iPSCs and mESCs, respectively, over K562 levels. However, the frequency of these mutations remained low, and not reproducible (Figure S15), which also explains the increase in overall between-replicate divergence (Figure 5B). Deletions attributed to microhomology-mediated end joining were 40% less frequent in RPE-1 samples compared to K562; instead these cells displayed more than double the rate of single base insertion (37% in RPE-1 vs 14% in K562). The same bias

Europe PMC Funders Author Manuscripts

was present in CHO cells, whereas both iPSC and mouse stem cells favored microhomology-mediated deletions at the expense of single base insertions (Figure 5A). This trend was recapitulated in the mutation class of the dominant mutation for each gRNA, which changed depending on the cell line (Figure 5D). We found little influence of the genetic background on the link between microhomology and repair outcomes (Figure 5C), replicating our findings in K562 in other lines and species.

Multiple Cas9 effector proteins with augmented properties have been engineered, which could also give rise to changes in observed mutations. We thus considered alternative CRISPR/Cas9 reagents in K562 cells, both enhanced Cas9 eSpCas9(1.1) (eCas9)[21]) and Cas9 fused to Three Prime Repair Exonuclease 2 (TREX2), which is known to increase deletion size[13,22]. eCas9 behaved similarly to Cas9 (Figure 5A and B), albeit with a slower editing saturation (Figure S2), whereas outcomes in the Cas9-TREX2 fusion protein line were markedly different from the others (Figure 5B). Cas9-TREX2 mutations were shifted towards larger deletions (Figure 5D and E), and favoured ligation of the intact PAM-proximal side with a deletion on the PAM-distal side (Figure S16), at the expense of frequent microhomology-mediated deletions (Figure 5C, S17). The generation of less biased and larger deletions, as generated by this fusion, could be beneficial in some contexts. We also tested a Cas9-2A-TREX2 construct harboring a 2A linker peptide which results in equal expression of monomeric Cas9 and TREX2. This construct did not generate additional larger deletions as observed for Cas9-TREX2 fusion, but did increase the frequency of single base deletions while reducing microhomology-mediated ones (Figure S17, S18). Combined, these data are consistent with the function of TREX2 as an exonuclease[23] that promotes repair via the canonical non-homologous end joining pathway, resulting in small deletions that are not mediated by microhomology[24,25].

Repair outcomes can be accurately predicted

So far, we have demonstrated that the repair outcomes are reproducible, biased, dependent on the local sequence, and mostly consistent across genetic backgrounds. These observations suggest that mutations generated by Cas9 ought to be predictable from sequence alone. To test this hypothesis, we developed a computational predictor of the mutational outcomes of a given gRNA, which we call FORECasT (Favoured Outcomes of Repair Events at Cas9 Targets). To accomplish this, we first generated candidate mutations for each gRNA, and derived features for them based on local sequence characteristics (Online Methods). We then split the set of available gRNAs into training, validation, and test sets, and trained a multi-class logistic regression model that minimizes the average KL divergence between predicted and actual repair profiles (Figure 6A).

The theoretical prediction accuracy limit is measurement repeatability. FORECasT achieves performance close to this limit, with the average KL between predicted and measured profiles only a little higher than between measurements in biological replicates (0.68 vs 0.65; Figure 6B; examples of predictions at the quartiles and whiskers in Figure S19). Frequencies of individual mutations that were reproducible (I1, D1, D2 and D>2 with microhomology) were also correspondingly well predicted (Figure S20). As a consequence, we could accurately predict the percentage of mutations that do not disrupt the reading

Europe PMC Funders Author Manuscripts

frame on held-out validation data (Pearson's R=0.81 for prediction vs 0.89 for replicates; Figure 6C, 2D). Despite being trained on K562 cells, FORECasT also achieves good accuracy on other cell lines (median KL range from 0.79 in CHO cells to 1.25 in mouse ES cells; Figure S21), with the magnitude of discrepancies consistent with shifts in the repair profiles (Figure 5A, 5B).

The sequence features with the largest weights mirror those that we observed to induce a bias in the outcomes, for example linking high frequency single nucleotide insertions to a repeat of a PAM-distal T nucleotide (Table S3). Individual deletion-related features (over 2000 total) had lower weights, most likely due to their larger quantity. Substantial biases in feature weights highlight the expected microhomology-related properties explored above, amongst others, and further experiments may yet elucidate additional sequence characteristics that promote particular repair outcomes.

Finally, we tested the extent to which the predicted rates of in frame mutation explain the variability in gRNA efficacy seen in existing gene knock-out experiments and large scale screens. We predicted mutational outcomes for gRNAs targeting 10 genes[26], but while the estimated fraction of frameshift edits is concordant with the measurements where available (12 gRNAs; Figure 6C, S22, Table S4), it does not explain the observed phenotypic variation (Figure S23). We also calculated correlations between the predicted fraction of out-of-frame mutations, and gRNA efficacy for essential gene gRNAs in three large scale screening datasets (Online Methods), and found a significant, albeit small, link (Figure S24). The strength of the association increased with library design iterations and quality, suggesting that ability to generate frameshift mutations is an increasingly important consideration once other sources of variability have been accounted for. In agreement with[26], we further observed a weak but consistent association between the predicted fraction of out-of-frame mutations and phenotypic effect for gRNA that target outside protein domains (Figure S25), indicating that it is more important to disrupt the reading frame in those regions.

These results demonstrate that repair outcomes can be predicted from sequence alone, and in a manner that is expected to generalise to all sites in the genome. We have made the predictor available as a webtool at https://partslab.sanger.ac.uk/FORECasT, and as a command line tool on GitHub at https://github.com/felicityallen/SelfTarget.

Discussion

We have presented the—to our knowledge—most comprehensive study of DNA double strand break repair outcomes to date. The Cas9-generated alleles show strong sequence dependent biases that are reproducible and predictable for dominant categories of mutation (single base insertions, small deletions, and microhomology-mediated deletions), despite some variability between genetic backgrounds and species.

Stem cells (human iPSCs and mouse ESCs) had a higher rate of large insertions than other lines, and favored microhomology-mediated deletions. These biases likely reflect different absolute and relative activities of the various DNA repair mechanisms. Preference for microhomology-mediated repair in stem cells may be linked to increased rates of homology-

Europe PMC Funders Author Manuscripts

directed repair, which shares the initial resection step[4], whereas favoring of single based insertions in CHO and RPE-1 lines indicate elevated canonical end-joining activity. The higher incidence of large insertions in stem cells could similarly be explained by aberrant homology-directed repair, where strand invasion occurs in the wrong place, such that DNA synthesis before strand displacement leads to additional sequence.

The strong sequence biases observed here for single nucleotide insertions, as previously also seen in yeast, and to some extent humans[9] could be explained by a model where the Cas9 protein stays bound to the PAM-proximal side of the cut while the staggered one-nucleotide overhang on the PAM-distal side[10,17] is filled by DNA polymerase, and re-ligated via the non-homologous end joining pathway[9,28]. Favoring of thymine insertion by this event could indicate either a preference of the DNA repair enzymes (especially polymerases), difference in availability of the required nucleotide triphosphate for incorporation, or propensity of Cas9 to make a staggered, rather than blunt cut when thymine is present. Finally, removal of one or two repeating nucleotides is the most frequent deletion, which could be achieved by processing and re-ligating the ends at the cut via a similar staggered intermediate.

Our assay was limited to confident detection of deletions of at most 30 base pairs, as the longer deletions would also remove the unique sequence we use to assign the measurement to a gRNA-target pair. Recent reports indicate that substantially larger events happen at non-negligible frequency[18], and indeed, some such outcomes explain the large discrepancies observed for a small number of gRNAs between our measurements and endogenous profiles. Nevertheless, in 94% of cases measured, there is good agreement between the outcomes we measured in our synthetic targets, and those at genomic targets.

Many genetic diseases, like Huntington's or Fragile X syndrome, are due to expansions of short tandem repeats[29]. Such repetitive sequence serves as excellent substrate for microhomology-mediated repair, and potential correction using Cas9, especially if they also harbor a PAM site, like the CGG expansion of the *FMR2* gene in fragile XE syndrome[30]. In the future, contraction of such rogue expansions could be explored as a therapy option, as the low efficacy allele replacement is not required, and simply generating a double strand break would shorten the pathogenic repeat. Indeed, a few preliminary efforts in this direction have already given promising results[31–33], but given the possible unintentional genomic damage[18], utmost rigor is required to demonstrate safety before any applications in humans. The data and model presented here will help in guiding gRNA design towards the desired outcomes for genome-wide screens and bespoke edits.

Online Methods

The code for all analyses is available at https://github.com/felicityallen/SelfTarget, as a runnable capsule on www.codeocean.com, and as Manuscript Related File 1.

Selection of guides and targets

We compiled our library of 41,630 total gRNA-target pairs (Dataset 1) from five sub-libraries aimed at testing different aspects of repair outcome generation:

Europe PMC Funders Author Manuscripts

1. Endogenous gRNA-Targets: We included 86 gRNAs used in[7] ("Endogenous targets") that were compatible with our library cloning method (see below), of which 9 were lost during cloning, giving 77 in total.

2. Genomic gRNA-Targets: We selected 6,568 gRNAs from existing gRNA libraries or literature, which we expect to have sequence characteristics representative of gRNAs in practical use. These included the Endogenous set above, as well as 5,431 from the Human v1.0 library[34]. Of these, 5,194 were selected because they were also present in the library used in[13,35] - these guides were obtained by filtering for common guides between the two libraries, and then discarding those that were incompatible with our assay (see below). The result was a set targeting 5,192 different human genes. Further 903 guide-target pairs were included from within the set used in[13] (again filtered for assay compatibility), which target genes that were considered by the authors to be of high value, as they targeted ion channels, receptors and genes in the cancer gene census. The remaining 234 gRNAs were designed with other in-house experiments in mind and targeted a range of essential and non-essential genes, or were in the Endogenous set above.

3. Explorative gRNA-Targets: We designed 27,906 guides to cover a wide range of local sequence characteristics. This included gRNAs with varying stretches, distances, and nucleotide compositions of microhomologous sequences as described below. The target sequences were randomly generated (except for the PAM) and then adjusted iteratively until the desired microhomology properties were achieved, ranging from targets with no microhomologous sequences longer than 2 nucleotides within 20 nucleotides of the cut site, to targets with microhomologous sequences up to 15 nucleotides long at close range. A larger set of gRNA-targets was initially created and then filtered both for compatibility with cloning (see below), as well as to ensure each gRNA had no direct targets in the human genome (>1 mismatch between gRNA and target). The sequences of the resulting set are in Dataset 1, and an overview is given in Table S5.

4. Microhomology Mismatch gRNA-Targets: 571 gRNA-targets were randomly selected from the Explorative gRNA-Targets above that had microhomology span lengths of 6 and above. These were randomly altered to change one (284 gRNAs) or two bases (287 gRNAs) in the microhomologous sequence.

5. Conventional Scaffold gRNA-Targets: All of the above subsets used the improved gRNA scaffold[17]. 6,218 gRNA-target pairs, all of which were already included in one of the first three subsets above (77 Endogenous, 3,777 Genomic (distinct set from 3,777 used for across cell line comparisons), 2364 Explorative), were ordered as separate oligos in the purchased oligo pool and were independently cloned with the alternative conventional gRNA scaffold[16]. These gRNAs allowed assessment of the impact of a difference in gRNA scaffold (which appears very small, Figure 2C) as well as providing independently synthesized and constructed repeat measurements of the same gRNAs.

Every target is uniquely barcoded by 10nt sequence both in the 3' and the 5' end (at least two mismatches between any two barcodes, randomly generated), to allow identification of each construct even in the absence of the full targeted sequence. All constructs passed the

Europe PMC Funders Author Manuscripts

filters of having no stretches of five adjacent nucleotides with at least four thymines in the gRNA sequence, since this can cause early termination of transcription; carrying no BbsI restriction sites or common primer sequences in the gRNA sequence or context, and not cutting elsewhere in the plasmid. All constructs were altered to contain a 'G' in the gRNA (but not the target) in the position 20nt before the PAM, for improved expression of the gRNA from the hU6 promoter.

Construction of lentiviral library

A lentiviral gRNA expression vector lacking the scaffold pKLV2-U6(BbsI)-PKGpuro2ABFP-W was generated by removing the improved gRNA scaffold from pKLV2-U6gRNA5(BbsI)-PKGpuro2ABFP-W[34] (Addgene #67974). This strategy allowed us to clone gRNA-target libraries encoding gRNAs linked to either the conventional or the improved scaffold sequences, but otherwise identical.

We generated by PCR on pKLV2-U6gRNA5(BbsI)-PKGpuro2ABFP-W two fragments encompassing the 5' end of the AmpR cassette to U6 promoter (primers P1-P2, Table S6) and PGK promoter to the 3' end of the AmpR cassette (primers P3-P4), respectively. Primer overhangs were designed to generate overlapping ends and pKLV2-U6(BbsI)-PKGpuro2ABFP-W was obtained by Gibson assembly (NEBuilder HiFi DNA Assembly Master Mix, NEB) of the two fragments. BbsI restriction sites were present downstream from the U6 promoter for subsequent cloning of the gRNA-target library inserts.

Library cloning started by PCR amplification of the 170 nt oligonucleotide pool of designed sequences (CustomArray) encoding gRNA and target sequence, separated by a spacer harbouring two BbsI restriction sites (Supplementary Figure 1), enclosed by priming sites, and using all the remaining 79nt to randomize the sequence context of the target. Primer pairs P5-P6 and P7-P8 (Table S6) were used to amplify oligos compatible with the conventional or improved scaffold respectively. Gibson assembly[36] was employed to fuse the amplified pool to a 193 nt G-block fragment (IDT) encoding either a conventional or improved version of the gRNA scaffold and a spacer. 1:1 molar ratio was mixed in three reactions incubated 1 h at 50ºC and subsequently pooled. The resulting 318 bp circular DNA was column purified (PCR purification kit, QIAGEN) and treated with Plasmid-Safe ATP-Dependent DNase (Epicentre) to remove linear DNA, followed by linearisation with BbsI at 37ºC for 2 h. The resulting 296 bp linear fragment was ligated into scaffold-less pKLV2-U6(BbsI)-PKGpuro2ABFP-W. Ligations (T4 DNA Ligase, NEB) were performed in triplicate, pooled and used in up to 10 electroporation reactions to maximise library complexity.

Generating the TREX2 construct

The Cas9-TREX2 and Cas9-2A-TREX2 vectors were made by fusing the human TREX2 open reading frame (GBlock, IDT) to the C-terminus of the Cas9 protein[37] with a GGGS linker or an intervening T2A peptide. These were cloned by Gibson assembly into a piggyBac vector (pKLV-Cas9), driven by a EFS promoter, and containing a blasticidin selectable marker. Genbank files of the final vectors are provided in the Supplementary Information.

Europe PMC Funders Author Manuscripts

Cell culture

K562, K562-Cas9 (kind gift by Etienne De Braekeleer) and all K562-derived lines (see below) were cultured in RPMI supplemented with 10% FCS, 2 mM L-glutamine, 100 U/ml penicillin and 100 mg/ml streptomycin. CHO-Cas9 and 293FT (Invitrogen) cells were cultured in Advanced DMEM supplemented with 10% FCS, 2 mM L-glutamine, 100 U/ml penicillin and 100 mg/ml streptomycin. HAP1-Cas9 were cultured in IMDM supplemented with 10% FCS, 2 mM L-glutamine, 100 U/ml penicillin and 100 mg/ml streptomycin. RPE-1-Cas9 cells were cultured in DMEM:F12 supplemented with 10% FCS, 0.26% sodium bicarbonate, 2 mM L-glutamine, 100 U/ml penicillin and 100 mg/ml streptomycin.

E14TG2a mouse ES cells supplied by Dr Meng Li (Cambridge Stem Cell Institute) were cultured in High glucose DMEM supplemented with 15% FBS, 2 mM L-glutamine, 0.1 mM 2-mercaptoethanol and 1,000 U/ml leukemia inhibitory factor (LIF; Millipore).

iPSCs (REC 15/WM/0276) were cultured in vitronectin (Life Technologies Ltd.) coated plates and TeSR-E8 medium (Stemcell Technologies). E8 media was changed daily throughout expansion and all experiments. All cell lines were cultured at 37ºC, 5% CO_2.

Lentivirus production and transduction of cell lines

Supernatants containing lentiviral particles were produced by transient transfection of 293FT cells using Lipofectamine LTX (Invitrogen). 5.4 μg of a lentiviral vector, 5.4 μg of psPax2 (Addgene #12260), 1.2 μg of pMD2.G (Addgene #12259) and 12 μl of PLUS reagent were added to 3 ml of OPTI-MEM and incubated for 5 min at room temperature. 36 μl of the LTX reagent was then added to this mixture and further incubated for 30 min at room temperature. The transfection complex was added to 80% confluent 293FT cells in a 10-cm dish containing 10 ml of culture medium. After 48 h viral supernatant was harvested and stored at –80 °C. Fresh medium was added and lentiviral supernatant was collected a second time 24 h later. When necessary we prepared larger amounts of lentivirus by scaling up the procedure above.

For lentiviral transduction K562 and K562-derived cells (see below), CHO-Cas9, HAP1-Cas9 and RPE-1-Cas9 cell lines were incubated with the lentiviral supernatant in a single cell suspension in presence of 8 μg/ml polybrene (Hexadimethrine bromide, Sigma) followed by centrifugation for 30 min at 1,000xg. E14TG2a mouse ESCs transduction was performed incubating cells in suspension for 30 min in presence of 8μg/ml polybrene. iPSC transduction was performed in a single cell suspension obtained by incubating cells with Accutase for 10 min (Millipore Corporation) and cells were plated in E8 medium supplemented with 8μg/ml polybrene.

Generation of K562-eCas9, K562-Cas9-TREX2, K562-Cas9-2A-TREX2 and E14TG2a-Cas9 lines

Cell lines stably expressing Cas9 were generated by lentiviral transduction followed by selection in the presence of Blasticidin (Cambridge Bioscience) to ensure high Cas9 activity. K562 cells were transduced using eSpCas9(1.1)[21] (Addgene #71814), Cas9-TREX2 and Cas9-2A-TREX2 vectors to generate K562-eCas9, K562-Cas9-TREX2 and K562-Cas9-2A-

Europe PMC Funders Author Manuscripts

TREX2 respectively. E14TG2a-Cas9 cells were generated by transducing E14TG2a cells with pKLV2-EF1aBsd2ACas9-W[34](Addgene #67978).

Screening and sequencing of repair outcomes

Cell lines were infected at a multiplicity of infection (MOI) ranging from 0.5 to 0.6 and at a coverage ranging from 500X to 1600X. Total number of cells, MOI and coverage for each screen are listed in Table S1. For each line, at least two separate infections were performed and treated separately as biological replicates. 24 h after transduction (72 h for iPSCs) puromycin was applied to the culture medium to select for successfully transduced cells and maintained throughout the screen. Cells were cultured for 7 days post infection, with a small number of samples further maintained for up to 20 days to evaluate the effect of timepoint choice (Table S1). Enough cells were passaged and collected to maintain coverage higher than at the time of infection.

Upon collection cells were centrifuged and pellets were stored at -20ºC prior extraction of genomic DNA. Briefly, cell pellets were resuspended into 100 mM Tris-HCl pH 8.0, 5 mM EDTA, 200 mM Nacl, 0.2% SDS and 1 mg/ml Proteinase K and incubated at 55ºC for 16 h. DNA was extracted by adding 1 volume of isopropanol followed by spooling, washed twice in 70% ethanol, centrifuged and resuspended into TE.

For sequencing, the region containing the target surrounded by the context was amplified by PCR using primers P10-P12 or P11-P12 respectively for the conventional and improved scaffold (Table S6) with Q5 Hot Start High-Fidelity 2× Master Mix (NEB) with the following conditions: 98 °C for 30 s, 24 cycles of 98 °C for 10 s, 61 °C for 15 s and 72 °C for 20 s, and the final extension, 72 °C for 2 min. Alternatively, both gRNA and target were amplified using primers P9-P12. For each gDNA sample the amount of input template was calculated keeping into account coverage and the amount of gDNA per single cell depending on the species and the ploidy of each line, and PCR reactions were scaled up accordingly. The PCR products were pooled in each group and purified using QIAquick PCR Purification Kit (QIAGEN). Sequencing adaptors were added by PCR enrichment of 1 ng of the purified PCR products using forward primer P13 and indexing reverse primer P14 with KAPA HiFi HotStart ReadyMix with the following conditions: 98 °C for 30 s, 12-16 cycles of 98 °C for 10 s, 66 °C for 15 s and 72 °C for 20 s, and the final extension, 72 °C for 5 min. The PCR products were purified with Agencourt AMPure XP beads, quantified and sequenced on Illumina HiSeq2500 or HiSeq4000 by 75-bp paired-end sequencing using the following custom primers: P15-P18 for sequencing of both gRNA and target, P16-P18 (conventional scaffold) or P17-P18 (improved scaffold) for target-only sequencing.

Sequence analysis

We processed the generated sequence data to compile repair profiles as follows. First, we combined the partially overlapping paired-end reads into a single sequence using pear v. 0.9.10[38] with options “-n 20 -p 0.1” (minimum combined sequence length of 20, probability of no overlap below 0.1). To assign reads to constructs, we required that at least one of the unique 3’ and 5’ barcodes be present with at most one mutation, and confirm that the read can be aligned to the template in such a way that at least 80% of the read characters has a

Europe PMC Funders Author Manuscripts

match in the construct template (i.e. there can be a large deletion in the read around the cut site, but minimal misalignment outside that region). The alignment was done using a custom dynamic programming algorithm in which the two sides of the cut site are independently aligned and then efficiently combined. This algorithm allows large deletions at a specified place (the expected Cas9 cut site) without penalty, while imposing substantial gap penalties elsewhere, and unlike generic tools, works for relatively short sequences. Once reads were assigned to oligos we checked each sample to ensure that the per-oligo log2 read counts (including those both with and without indels) of the "Explorative gRNA-Target" set (since these have no direct targets in the genome) were well correlated (Pearon's $R > 0.95$, computed using scipy.stats.pearsonr[39]) with those in the original plasmid library to minimize distortion in the measured mutational profiles that could be due to reasons other than Cas9-cutting and subsequent cellular repair.

The mapping and alignment of sequences was first carried out on the plasmid library to compile a set of null mutations that are present before any editing experiments. These null mutations were then used as templates for alignment of sequencing reads from the screens, again using the custom dynamic program, so that mutations already existing in the plasmid library (due to oligo synthesis errors or somatic mutation) are not erroneously attributed to Cas9 activity. Additionally, a deeper measurement was taken of the library in K562 cells without Cas9 present, processed as for all other samples, and then used to filter the other profiles to remove all mutations seen in these non-Cas9 samples unless they were present with at least three times their non-Cas9 frequency in the other sample (since, for example, some very low-frequency technical artifacts can resemble microhomology-mediated deletions).

Adequate coverage was ensured by only including gRNAs in the sub libraries described above if they had at least 20 mutated reads in all three K562 replicates. For analysis involving other cell lines, this criterion was extended to ensure that each gRNA had at least 20 mutated reads in all samples employed. For example, the analysis of microhomology effect in non-K562 cells used a restricted subset of 16,272 gRNA-target pairs from the explorative gRNA target set, and the remaining results presented in Figure 5 used 3,777 gRNA-Targets from the Genomic gRNA-Targets set, all restricted to those with over 20 mutated reads in all K562 and non-K562 samples.

Note that although a proportion of the gRNAs used target essential genes (178 target genes in Hart's essential gene set[40]), and may therefore be expected to decrease in coverage as the assay goes on, the limited duration of the assay (7 days), and the fact that the measurement is made at an independent synthetic target rather than at the endogenous gene, is expected to result in negligible impact on the mutational profiles measured. Indeed, assuming independence of editing events, the only expected effect of fitness differences should be on coverage, which we account for by ensuring adequate read counts as above. Correspondingly, we observed no bias in KL values for the 189 guides targeting essential genes (as defined by[40]) compared to those that do not (Figure S26).

Europe PMC Funders Author Manuscripts

Repair profile comparisons

We store the repair profile as a collection of read counts per indel, where each indel is characterised by its size, type and location with respect to the cut site, as specified by an identifier string e.g. the identifier 'D2_L-3R0' describes a size 2 deletion for which the last unaltered nucleotides are 3 to the left of the cut site, and at the cut site, respectively. To calculate similarity of repair profiles, we use the symmetrized Kullback-Leibler divergence ("KL"). Standard Kullback-Leibler divergence is calculated as

$$D_{KL}(P\|Q) = \Sigma_i \; P_i \, log \frac{P_i}{Q_i},$$

where i indexes the different indels, and P_i and Q_i are their normalized proportion of total mutated read counts in the compared samples; we employ the symmetric form,

$$KL = D_{KL}(P\|Q) + D_{KL}(Q\|P).$$

To avoid division by zero, we add small pseudocounts of 0.5 to all indels present in one sample, but not the other, in the computation. To give the reader a sense of the similarity of profiles with various KL values at the respective quantiles and whiskers of our measurements in Figure 6B, we provide a series of examples in Figure S19(a-d).

For frame-shift comparisons, we accumulated all reads for all mutations for a given gRNA-target such that those mutations whose size are a multiple of 3 are considered "in frame" and those whose size are not a multiple of 3 are considered "out of frame". Unmutated reads were discarded from all comparisons (except read counts and mutation rates presented in Figures S2 and S5).

Repair profile analysis

We classified all indels into types I1 (size 1 insertion), D1 (size 1 deletion), D2 (size 2 deletion I>1 (larger insertions), I+D (insertion and deletion), and D>2 (larger deletions). For deletions of size greater than 2 we assigned a likely causal mechanism of the event as likely generated by microhomology if at least 2nt of matching sequence were present on either side of the cut, and likely not generated by microhomology otherwise. To analyse indel prevalence by size (Figure 3A), we classified indels into deletions of 1-30 nt and insertions of 1 to 10 nt. In this case, information about deletions larger than 30 and insertions larger that 10 nt was excluded from the analysis since they are not well-detected by our method. In this, and other results presenting accumulated measurements across gRNAs (Figure 3A,3B, 3C,5A) we first normalized all indel counts for each gRNA by the total number of mutated reads for that gRNA, such that all gRNAs weigh equally towards each measurement, rather than proportionally to their read coverage.

Predicting repair profiles

For each gRNA-target pair we generated candidate indels by considering all possible insertions up to size 2, within 3 nucleotides of the cut site, and all deletions spanning the cut

Europe PMC Funders Author Manuscripts

site with a left edge from one right of the cut site to up to 30 nucleotides left of the cut site and a right edge up to 30 nucleotides the other way, up to a maximum deletion size of 30. For each of these candidates, we computed a set of 3,633 binary features that describe the length, location and nucleotide composition of inserted sequences, microhomologies and their neighboring nucleotides. Many of these features also comprise pairwise 'AND' results between features to capture interaction effects.

We modelled the probability of each possible outcome using a logistic that ensures the sum of all possible outcomes for a give gRNA sums to one. i.e. the probability of the j-th mutation for a given gRNA is modelled as

$$p_j = exp\left(\theta x_j\right) / \sum_i exp\left(\theta x_i\right)$$

where θ is the vector of parameter weights, and x_j is the feature vector for that mutation; the sum is over all mutations for a particular gRNA. We then minimize the L2-regularized, non-symmetric KL divergence of these probabilities when compared to the measured proportions, by computing closed form partial gradients with respect to the kth parameter θ_k and using L-BFGS-B within scipy.optimize.minimize to perform gradient descent optimization of this metric[39,41].

For development of the predictor, we randomly selected gRNA-target pairs from the "Explorative gRNA Targets" set, restricted to those with more than 100 reads in K562 cells, and without a corresponding counterpart in the "Conventional scaffold gRNA Targets" set, and performed training and hyperparameter tuning by randomly selecting two disjoint sets of N=50,100,200,500,1000,5000 gRNAs from this set and assigning these to training and test sets respectively, and repeating this 3 times for each hyperparameter and training set size. With 5000 training and test examples, the training and test scores converged for this feature set with an L2 regularization constant of 0.01, so the parameters trained with these settings were selected for further validation (Figure S27). We used the "Conventional scaffold gRNA Targets" set as a held-out validation set, predicting profiles by applying the associated predicted probabilities to generate 1000 counts each for all gRNA-target pairs (dropping mutations predicted to have less than 1 count). We then validated the accuracy of these profiles by comparing against measurements for these gRNA-targets. Replicate A in Figure 6 (and Figure 2D) summed counts from both 800x DPI7 replicates from the K562 Improved scaffold samples, whereas Replicate B used the single replicate 1600x DPI7 sample.

Endogenous data processing

We collected the raw read data from the SRA archives referenced by[7] and[26] and re-processed the generated mutations for each gRNA using the same custom alignment program we used for our own data. For the Overbeek data, we used the data they collected in K562 at day 11 following lentiviral transduction, for closest compatibility with our own data, and accumulated reads across replicates into a single replicate for each gRNA. The

Europe PMC Funders Author Manuscripts

data for the heights and locations of the Shi *et al.* data used in Figure S23 was obtained via personal communication with the authors; we thank them for their assistance.

To analyse the influence of chromatin on the results in Figure 2E, we downloaded the ChromHMM[42] chromatin state assignments for K562 cell line as measured in ENCODE[43] from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgSegmentation/wgEncodeAwgSegmentationChromhmmK562.bed.gz, and used bedtools[44] to overlap them with the target locations in van Overbeek *et al.*

Frame-shift assessment in screen data

We tested for association between the predicted fraction of out-of-frame outcomes, and the gRNA efficacy of known essential genes in three large scale screening datasets; Meyers et al. 2017 (Avana library)[45], Tzelepis et al. 2016 (Yusa v1.0 library)[34], and Aguirre et al. 2016 (GeCKO v2 library)[46]. Our metric for relative gRNA efficacy was the gRNA scores inferred by JACKS[47], which provides a multiplier to the expected log-fold response of each gRNA compared to other gRNAs targeting the same gene. We used the JACKS inferred gRNA outputs available at https://figshare.com/articles/Results/6002438, and restricted the examined associations to gRNAs in these sets that target essential genes defined by Hart[40]. Pearson's R coefficients were calculated using scipy.stats.pearsonr[39].

To assess the importance of frame-shift rate for efficacy of gRNA targeting within or outside protein domains we mapped genomic location of the cut site to position in a protein for each gRNA targeting essential genes. We then found which cut sites are contained within protein domains. To do that, we used R package ensembldb and annotation package EnsDb.Hsapiens.v75 based on ENSEMBL version 75 and GRCh37 genome assembly[48].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

FA was supported by a Royal Commission for the exhibition of 1851 Research Fellowship. LP was supported by Wellcome (Wellcome 206194), and Estonian Research Council (IUT 34-4). HPH was supported Supported by a Wellcome Trust grant (Wellcome 200848/Z/16/Z) and a Wellcome Trust Strategic Award to the Cambridge Institute for Medical Research (Wellcome 100140). YG is funded by Cancer Research UK, C6/A18796 and a Wellcome Trust Investigator Award 206388/Z/17/Z in the Jackson lab. FMM was funded by a Marie Curie Intra-European Fellowship, project number 626375, DDR SYNVIA, and by a Wellcome Trust Investigator Award 206388/Z/17/Z and AstraZeneca Collaborative Award in the Jackson lab. We thank J. Eliasova for help with Figure 1, E. de Braekeleer from Wellcome Sanger Institute for providing the K562-Cas9 line, and A. Lawson for comments on the text.

References

1. Doudna JA, Charpentier E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science. 2014; 346
2. Chiruvella KK, Liang Z, Wilson TE. Repair of Double-Strand Breaks by End Joining. Cold Spring Harb Perspect Biol. 2013; 5
3. Her J, Bunting SF. How cells ensure correct repair of DNA double-strand breaks. J Biol Chem. 2018; 293:10502–10511. [PubMed: 29414795]

Europe PMC Funders Author Manuscripts

4. Truong LN, et al. Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. Proc Natl Acad Sci U S A. 2013; 110:7720–7725. [PubMed: 23610439]
5. Shibata A. Regulation of repair pathway choice at two-ended DNA double-strand breaks. Mutat Res. 2017; 803–805:51–55.
6. Bae S, Kweon J, Kim HS, Kim J-S. Microhomology-based choice of Cas9 nuclease target sites. Nat Methods. 2014; 11:705–706. [PubMed: 24972169]
7. van Overbeek M, et al. DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. Mol Cell. 2016; 63:633–646. [PubMed: 27499295]
8. 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. 2013; 32:267–273. [PubMed: 24535568]
9. Lemos BR, et al. CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. Proc Natl Acad Sci U S A. 2018; 115:E2040–E2047. [PubMed: 29440496]
10. Shou J, Li J, Liu Y, Wu Q. Precise and Predictable CRISPR Chromosomal Rearrangements Reveal Principles of Cas9-Mediated Nucleotide Insertion. Mol Cell. 2018; 71:498–509.e4. [PubMed: 30033371]
11. Taheri-Ghahfarokhi A, et al. Decoding non-random mutational signatures at Cas9 targeted sites. Nucleic Acids Res. 2018; doi: 10.1093/nar/gky653
12. Chakrabarti AM, et al. Target-specific precision of CRISPR-mediated genome editing. 2018; doi: 10.1101/387027
13. Chari R, Mali P, Moosburner M, Church GM. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nat Methods. 2015; 12:823–826. [PubMed: 26167643]
14. Kim HK, et al. In vivo high-throughput profiling of CRISPR–Cpf1 activity. Nat Methods. 2016; 14:153–159. [PubMed: 27992409]
15. Tycko J, et al. Pairwise library screen systematically interrogates Staphylococcus aureus Cas9 specificity in human cells. Nat Commun. 2018; 9:2962. [PubMed: 30054474]
16. Mali P, et al. RNA-guided human genome engineering via Cas9. Science. 2013; 339:823–826. [PubMed: 23287722]
17. Chen B, et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. Cell. 2013; 155:1479–1491. [PubMed: 24360272]
18. Kosicki M, Tomberg K, Bradley A. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. Nat Biotechnol. 2018; doi: 10.1038/nbt.4192
19. Cho SW, et al. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. Genome Res. 2014; 24:132–141. [PubMed: 24253446]
20. Gallagher DN, Haber JE. Repair of a Site-Specific DNA Cleavage: Old-School Lessons for Cas9-Mediated Gene Editing. ACS Chem Biol. 2018; 13:397–405. [PubMed: 29083855]
21. Slaymaker IM, et al. Rationally engineered Cas9 nucleases with improved specificity. Science. 2016; 351:84–88. [PubMed: 26628643]
22. Bothmer A, et al. Characterization of the interplay between DNA repair and CRISPR/Cas9-induced DNA lesions at an endogenous locus. Nat Commun. 2017; 8
23. Mazur DJ, Perrino FW. Excision of 3’ termini by the Trex1 and TREX2 3'-->5' exonucleases. Characterization of the recombinant proteins. J Biol Chem. 2001; 276:17022–17029. [PubMed: 11279105]
24. Bhargava R, Carson CR, Lee G, Stark JM. Contribution of canonical nonhomologous end joining to chromosomal rearrangements is enhanced by ATM kinase deficiency. Proc Natl Acad Sci U S A. 2017; 114:728–733. [PubMed: 28057860]
25. Certo MT, et al. Coupling endonucleases with DNA end-processing enzymes to drive gene disruption. Nat Methods. 2012; 9:973–975. [PubMed: 22941364]
26. Shi J, et al. Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nat Biotechnol. 2015; 33:661–667. [PubMed: 25961408]

Europe PMC Funders Author Manuscripts

27. Zuo Z, Liu J. Cas9-catalyzed DNA Cleavage Generates Staggered Ends: Evidence from Molecular Dynamics Simulations. Sci Rep. 2016; 5
28. Richardson CD, Ray GJ, DeWitt MA, Curie GL, Corn JE. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. Nat Biotechnol. 2016; 34:339–344. [PubMed: 26789497]
29. Sutherland GR, Richards RI. Simple tandem DNA repeats and human genetic disease. Proceedings of the National Academy of Sciences. 1995; 92:3636–3641.
30. Gu Y, Shen Y, Gibbs RA, Nelson DL. Identification of FMR2, a novel gene associated with the FRAXE CCG repeat and CpG island. Nat Genet. 1996; 13:109–113. [PubMed: 8673086]
31. Cinesi C, Aeschbach L, Yang B, Dion V. Contracting CAG/CTG repeats using the CRISPR-Cas9 nickase. Nat Commun. 2016; 7
32. Mahadevan MS, et al. Reversible model of RNA toxicity and cardiac conduction defects in myotonic dystrophy. Nat Genet. 2006; 38:1066–1070. [PubMed: 16878132]
33. Park C-Y, et al. Reversion of FMR1 Methylation and Silencing by Editing the Triplet Repeats in Fragile X iPSC-Derived Neurons. Cell Rep. 2015; 13:234–241. [PubMed: 26440889]
34. Tzelepis K, et al. A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. Cell Rep. 2016; 17:1193–1205. [PubMed: 27760321]
35. Wang T, Wei JJ, Sabatini DM, Lander ES. Genetic screens in human cells using the CRISPR-Cas9 system. Science. 2014; 343:80–84. [PubMed: 24336569]
36. Gibson DG. Enzymatic Assembly of Overlapping DNA Fragments. Methods in Enzymology. 2011:349–361. [PubMed: 21601685]
37. Cong L, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013; 339:819–823. [PubMed: 23287718]
38. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 2014; 30:614–620. [PubMed: 24142950]
39. JonesOliphantPeterson. SciPy: Open source scientific tools for Python. scipy; 2001. Available at: http://www.scipy.org [Accessed: 10th January 2018]
40. Hart T, et al. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. G3. 2017; 7:2719–2727. [PubMed: 28655737]
41. Zhu C, Byrd RH, Lu P, Nocedal J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Trans Math Softw. 1997; 23:550–560.
42. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012; 9:215–216. [PubMed: 22373907]
43. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
44. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014; 47:11.12.1–34.
45. Meyers RM, et al. Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet. 2017; 49:1779–1784. [PubMed: 29083409]
46. Aguirre AJ, et al. Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. Cancer Discov. 2016; 6:914–929. [PubMed: 27260156]
47. Allen F, et al. JACKS: joint analysis of CRISPR/Cas9 knock-out screens. bioRxiv. 2018; doi: 10.1101/285114
48. Zerbino DR, et al. Ensembl 2018. Nucleic Acids Res. 2018; 46:D754–D761. [PubMed: 29155950]

Europe PMC Funders Author Manuscripts

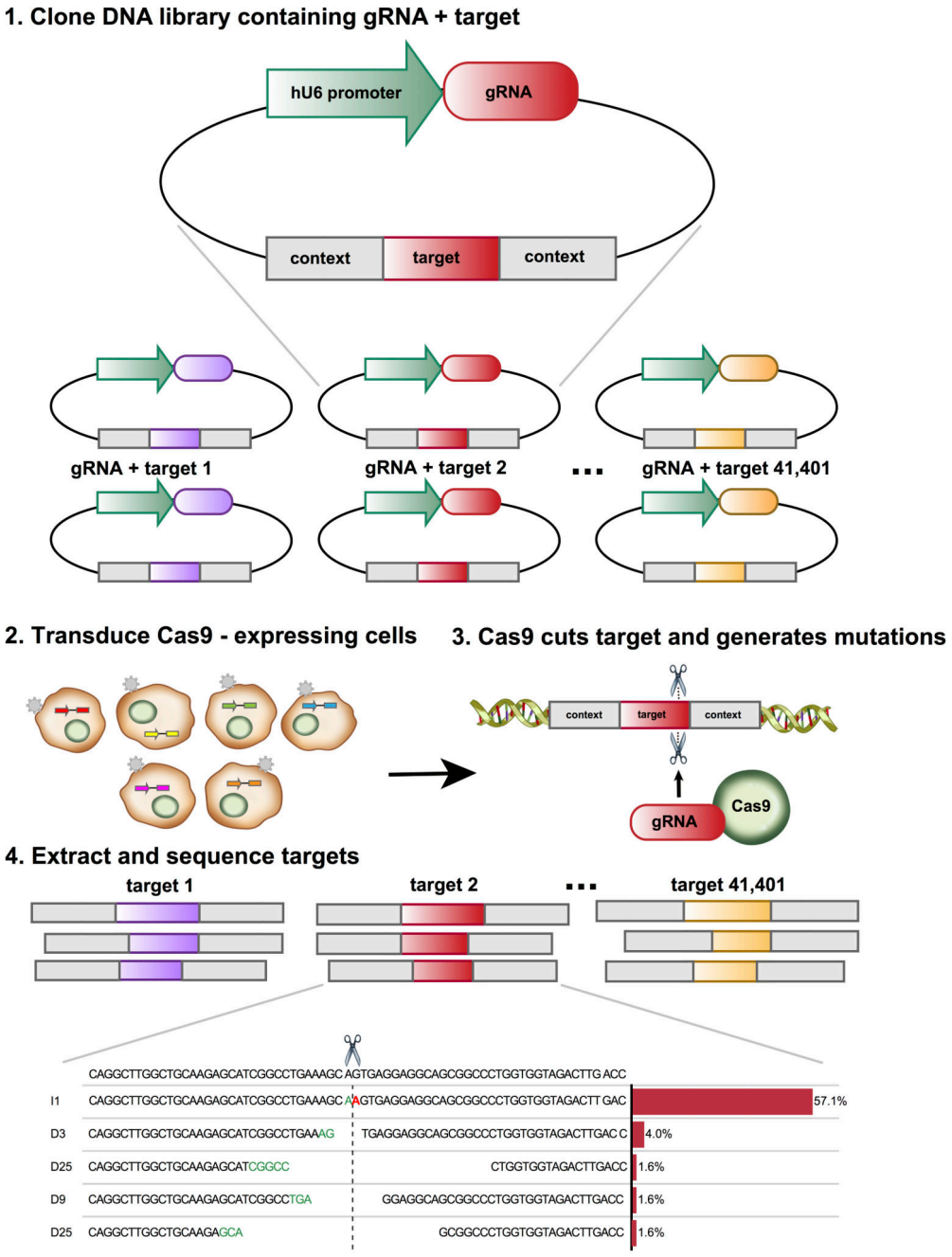

Figure 1. Mutational profiles generated by CRISPR/Cas9, and a method for their high-throughput measurement.

High throughput measurement of repair outcomes. Constructs containing both a gRNA and its target sequence (matched colors) in variable context (grey boxes) are cloned en masse into target vectors containing a human U6 promoter (green) (1), packaged into lentiviral particles, and used to infect cells (2), where they generate mutations at the target (3). DNA from the cells is extracted, the target sequence in its context amplified with common primers, and the repair outcomes (location, size, and sequence of mutation) determined by deep short read sequencing (4).

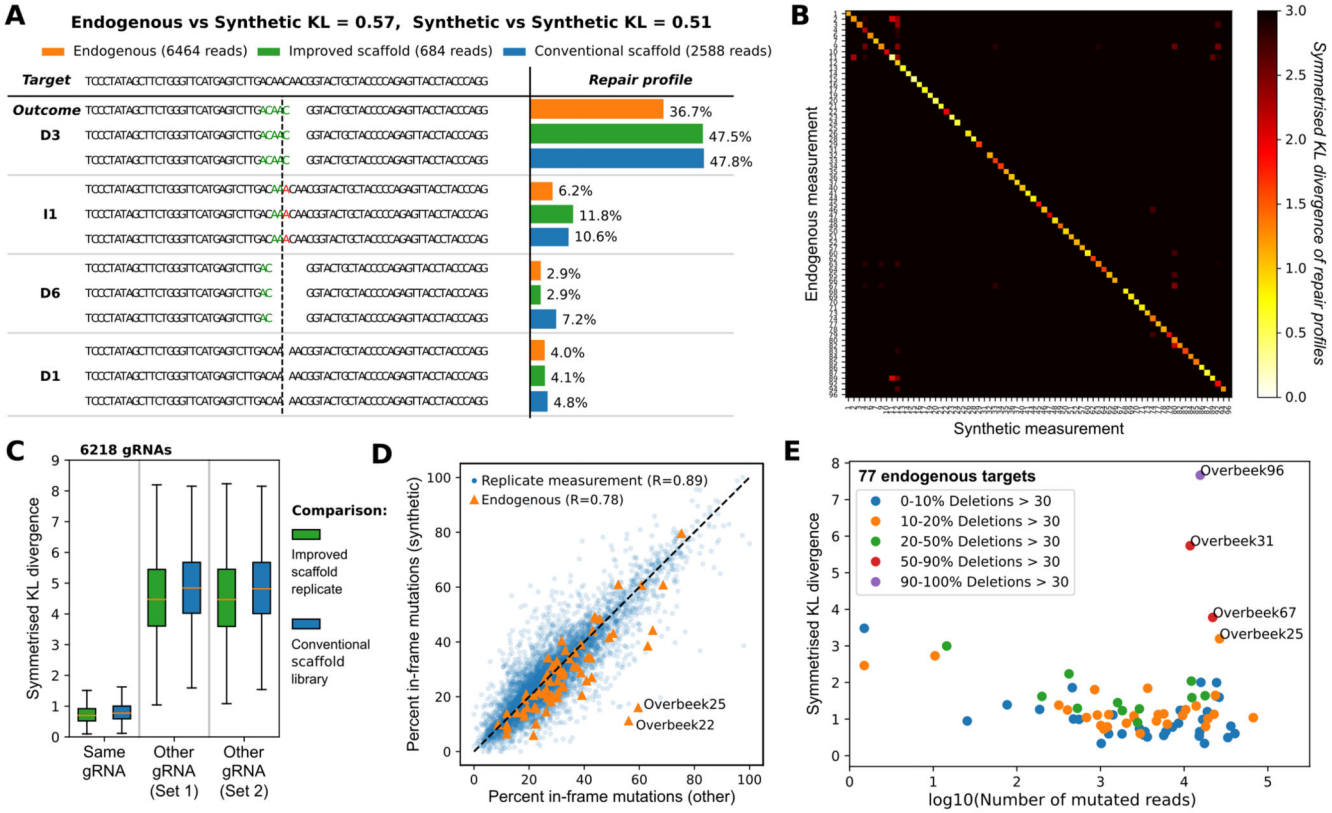

Figure 2. Synthetic mutational profiles are reproducible, specific to individual gRNAs and closely resemble endogenously measured profiles in human K562 cells.

A. Example of measured repair profile reproducibility for one gRNA-target pair. DNA sequence of the target (top) is edited to produce a range of synthetic outcomes that employ the improved gRNA scaffold (green bars) and conventional gRNA scaffold (blue bars), contrasted to endogenous measurements (orange bars). The proportions (x-axis) of the four most frequent mutational outcomes (e.g. “D3” - deletion of three base pairs depicted, “I1” - insertion of a single “A” at cut site, etc.; y-axis) is consistent between the experiments. Stretches of microhomology (green) and inserted sequences (red) are highlighted at the cut site (dashed vertical line).

B. Synthetic measurements faithfully capture endogenous outcomes. Symmetrized Kullback-Leibler divergence (white to black color scale) between synthetic repair profile measurements in K562 cells (x-axis) and endogenous repair profiles from van Overbeek et al. (y-axis; at least 100 reads in our synthetic samples).

C. Synthetic measurements are reproducible and gRNA-specific, irrespective of gRNA scaffold used. Box plots (orange median line, quartiles for box edges, 95% whiskers) of symmetrised KL divergences between two measurements of the same target (left), or between measurements of randomly selected target pairs from the same set (middle, right). Green boxes: comparison of biological replicates of the same library using the improved scaffold; blue boxes: comparison of matched measurements between libraries employing the conventional scaffold, and the improved scaffold; median mutated read numbers per gRNA

in parentheses. The 6,218 gRNAs used are from the “Conventional Scaffold gRNA-Targets” set (Online Methods); improved scaffold is used throughout the rest of the paper.

D. Frame information is reproducible between replicates, and well correlated with endogenous outcomes. Blue markers: Percentage of in-frame outcomes in our synthetic measurements (y-axis) contrasted against another biological replicate (x-axis; Pearson’s R=0.89, gRNAs as in C, improved scaffold only). Orange markers: same, but contrasting information from combined synthetic replicates (y-axis) against 68 endogenous measurements (x-axis; Pearson’s R=0.78, gRNAs as in B, excluding four with majority of large deletions not captured in our assay).

E. Low coverage and large deletions are the main sources of discrepancy between endogenous and synthetic measurements. Symmetrized KL divergence (y-axis) between endogenous and synthetic measurements of editing outcomes (individual markers; gRNAs as in B) is dependent on the sequencing coverage ($\log_{10}$(number of obtained reads), x-axis), and frequency of very large deletions (colors). Three target sequences that frequently give rise to very large deletions (red, purple) are not well captured by our assay design.

Europe PMC Funders Author Manuscripts

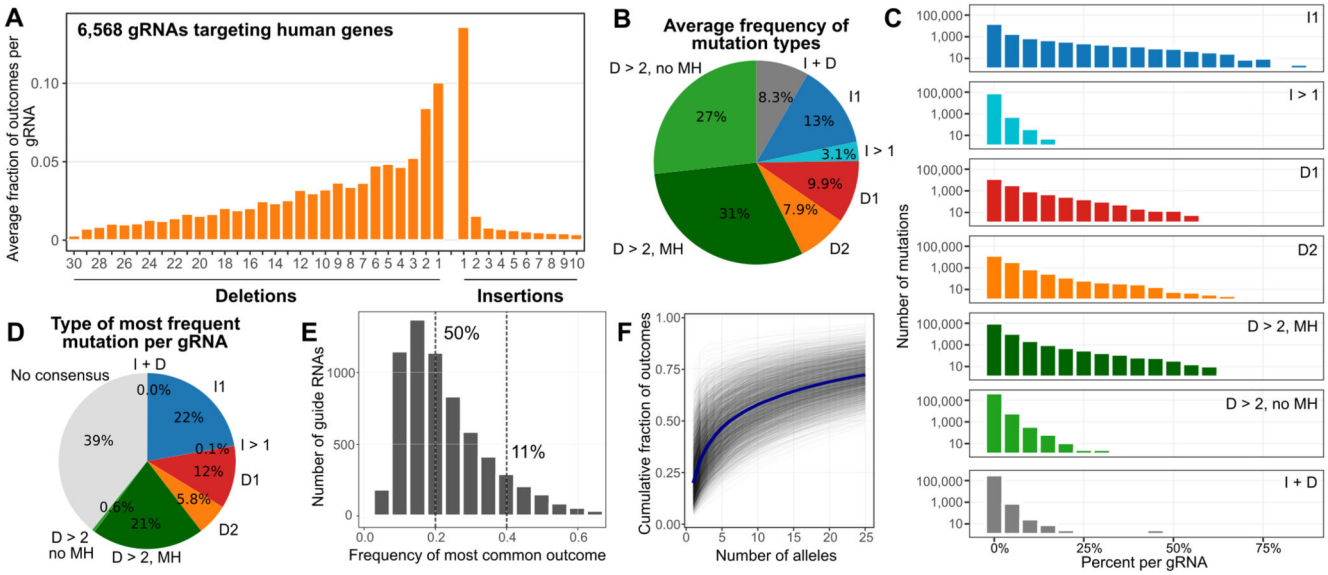

Figure 3. Mutational profiles are diverse and biased in K562 cells, as measured using 6,568 gRNAs with a median 991 sequenced reads with mutations per target.

A. Single base insertions are most common, with a long tail of moderately long deletions. The frequency (y-axis) of deletion or insertion size (x-axis), averaged across sequence targets present in the genome.

B. Editing outcome types are diverse. The percent occurrence per gRNA (area of wedge) of 1nt insertions (I1, blue), larger insertions (I > 1, teal), single base deletions (D1, red), dinucleotide deletions (D2, orange), larger deletions likely mediated by microhomology (D > 2, MH; dark green), other larger deletions (D > 2, no MH; light green), and more complex alleles (I + D, grey), measured in K562 cells, and averaged across genomic sequence targets.

C. Per-gRNA event frequencies differ across indel classes. Number of individual indels (y-axis, $\log_{10}$-scale) as a percentage of all mutations observed for their gRNA (x-axis) separated by mutation class (rows). Colors as in (B).

D. Specific single base insertions and microhomology-mediated deletions are the most frequent reproducible mutation classes. The percent of gRNAs (area of wedge) that have the same specific allele as their most frequent mutation in all three replicates, stratified by indel class (colors). 'No consensus': inconsistent most frequent mutation across replicates.

E. A single allele can account for a large fraction of editing outcomes for a gRNA. Number of gRNAs (y-axis) with the frequency of its most common outcome (x-axis) in K562 cells.

F. A small number of outcomes explains most of the observed data, but many low frequency alleles are present. Cumulative fraction of observed data (y-axis) matching an increasing number of outcomes (x-axis) for each target in K562 cells (grey lines), and their average (blue line).

Europe PMC Funders Author Manuscripts

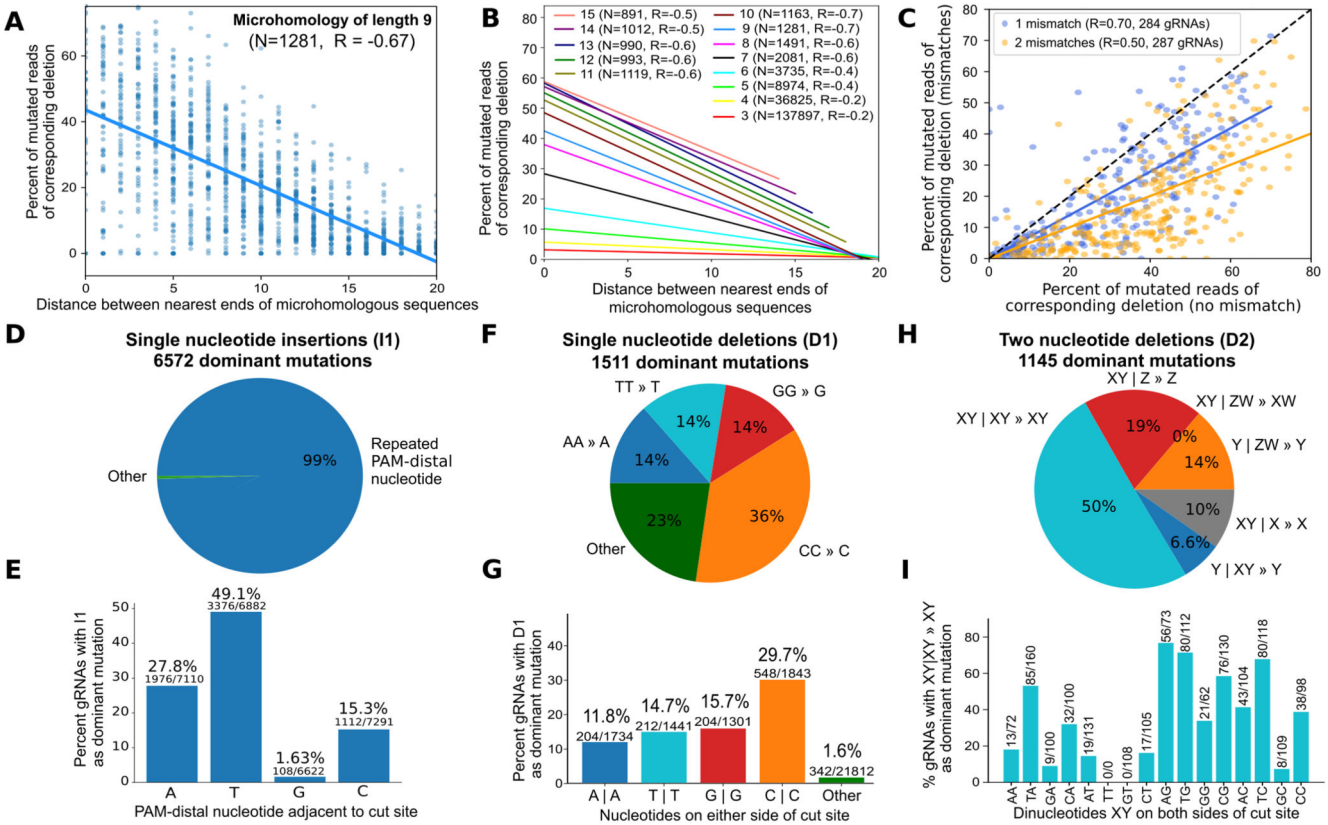

Figure 4. Local sequence context strongly influences editing outcomes in the explorative set of gRNA-target pairs.

A. Nearby matching sequences are used as substrate for microhomology-mediated repair more frequently than distant ones. Fraction of mutated reads (y-axis) for increasing distance between 1,281 matching sequences of length 9 (x-axis) (blue markers) in K562 cells, and a linear regression fit to the trend (solid line; Pearson's R=-0.67). Reproducibility of measurements is presented in Figure 5C.

B. Frequency of microhomology-mediated repair depends on the length of and distance between the matching sequences. Same as (A), but linear regression fits only for microhomologies of lengths 3 (red, bottom) to 15 (pink, top), with the number of pairs of matching sequences considered (N) and Pearson's correlation (R) noted in the legend.

C. Mutations in microhomology sequence reduce repair outcome frequency, but corresponding deletions are still present. The fraction of mutated reads associated with the particular microhomology with mismatches (y-axis) vs without mismatches (x-axis) stratified by the number of mismatches (blue: one mismatch, yellow: two mismatches). Solid lines: linear regression fits; dashed black line: y=x; Pearson's R provided in legend.

D. Single nucleotide insertions are only dominant when repeating the PAM-distal nucleotide. Percentage of the 6,572 gRNAs for which insertion of a specific nucleotide is most frequent in all replicates ("dominant allele"; area of wedge) stratified by whether the PAM-distal nucleotide adjacent to the cut site is inserted (blue), vs. all other outcomes (green).

Europe PMC Funders Author Manuscripts

E. Insertions of thymine dominate often, while guanines are rarely inserted with reproducibly high frequency. The percentage of gRNAs that have a dominant single nucleotide insertion (y-axis), stratified by their PAM-distal nucleotide at the cut site (x-axis).

F. Dominant single nucleotide deletions usually remove one nucleotide from a repeating pair at the cut site. Percentage of the 1,511 gRNAs with a dominant single nucleotide deletion (area of wedge) of a repeating A (blue), repeating T (teal), repeating G (red), repeating C (orange), or a base from a non-repeat (green).

G. Dominance of single nucleotide deletions depends on both bases adjacent to the cut site. The percentage of gRNAs that have a dominant single nucleotide deletion (y-axis), stratified by the two bases on either side of the cut site (x-axis).

H. Two nucleotide deletions that are dominant favour repeats. Percentage of the 1,145 gRNAs with a dominant size two deletion (area of wedge) that delete a repeat (XY | XY » XY, teal), delete PAM-distal nucleotides (XY | Z » Z, red), delete one PAM-distal and one PAM-proximal nucleotide (XY | ZW » XW, purple), delete PAM-proximal nucleotides (Y | ZW » Y, orange), delete a PAM-distal nucleotide flanked by a repeating base (XY | X » X, grey), or delete a PAM-proximal nucleotide flanked by a repeating base (Y | XY » Y, blue). X, Y, Z, W - any nucleotide; | - cut site.

I. PAM-distal guanine at the cut site promotes, while PAM-distal thymine at the cut site demotes the frequency of dominant dinucleotide repeat contraction. The percentage of gRNAs with a dinucleotide repeat that have the corresponding dominant two nucleotide deletion (y-axis), stratified by the two bases in the repeated sequence (x-axis).

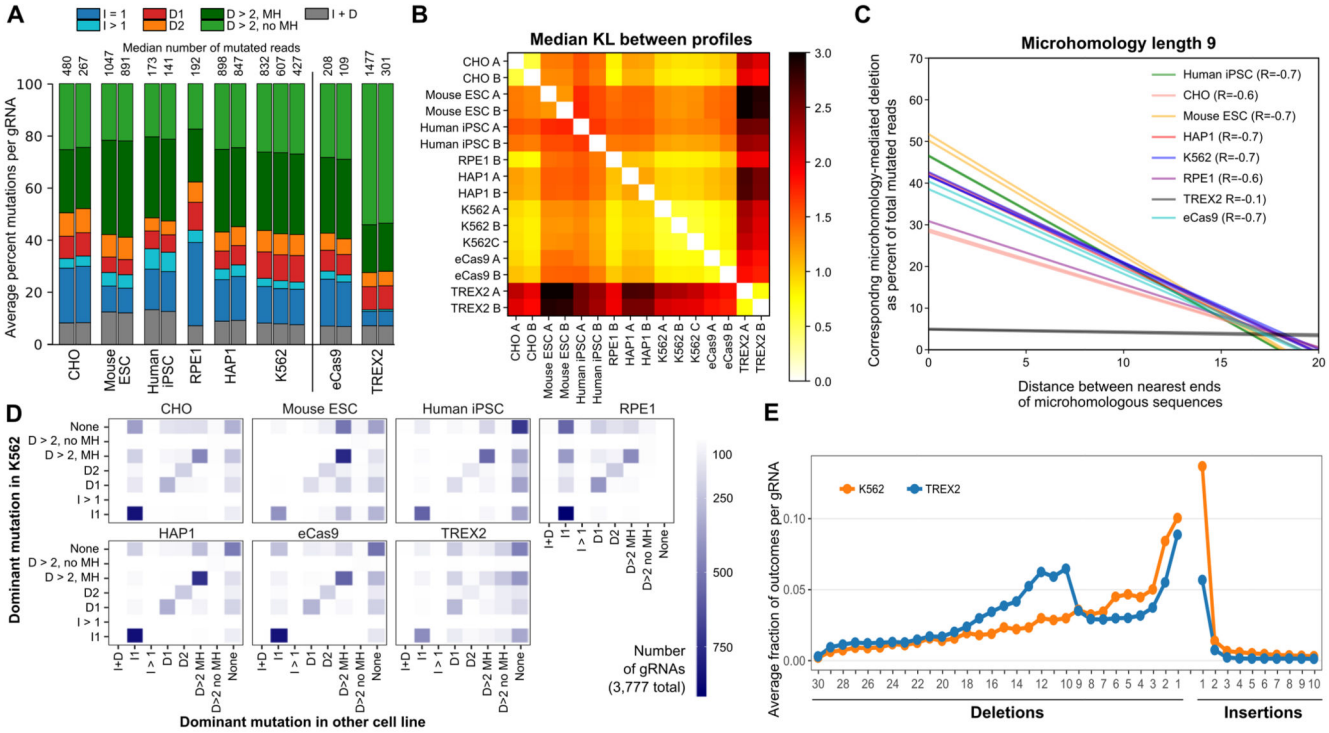

Figure 5. Differences between editing outcomes in K562-Cas9 and other cell lines and effector proteins.

A. Genetic background influences editing outcomes. Average per-gRNA frequency of different types of editing outcomes in 3,777 gRNAs (y-axis; colors as 3B) for Chinese hamster ovary cell line (CHO), mouse embryonic stem cells (Mouse ESC), human induced pluripotent stem cells (iPSCs), human retinal pigmented epithelial cells (RPE-1), human near-haploid cell line (HAP1), K562 cell line, and K562 cells with alternative Cas9 proteins: enhanced Cas9 (eCas9), and Cas9-TREX2 fusion (TREX2). Separate vertical bars are measurements from biological replicates; median number of mutated reads per gRNA is given above the bar for each replicate.

B. Mutational outcomes are similar across cell lines, with consistent moderate differences in stem cells and the K562 Cas9-TREX2 fusion line. Median symmetric Kullback Leibler divergence between repair profiles (black to white color range, as in Figure 2B) in different tested lines (x and y axis). gRNAs as in A.

C. Microhomology-mediated repair fidelity is similar across genetic backgrounds, but differs for Cas9-TREX2 fusion. Regression lines (as in Fig 4A) for fraction of mutated reads (y-axis) for increasing distance between matching sequences of length 9 (x-axis) in K562 cells (blue) and other tested lines (colors) in multiple replicates (individual lines), with overall Pearson's correlation denoted in the legend. gRNAs as in Figure 4B, restricted to those 822 gRNAs with MH of length 9 and at least 20 mutated reads in all samples.

D. The type of the dominant outcome per gRNA is consistent across cell lines overall, but biased towards microhomology-mediated deletions in stem cells, and I1 insertions in RPE-1 and CHO. The number of gRNAs (color) for which the most frequent indel comes from each class (x-axis) in the other cell lines examined (panels) compared to that for the same gRNA in K562 (y-axis). "None" refers to gRNAs without any indel consistently most frequent in

Europe PMC Funders Author Manuscripts

all replicates. gRNAs as in A. RPE data is based on one replicate, K562 on three, all other cell lines on two replicates.

E. Cas9-TREX2 fusion protein favours larger deletions compared to K562. Deletions of increasing size (x-axis) become more frequent (y-axis) in K562 Cas9-TREX2 cells (blue) compared to standard K562 Cas9 (orange). gRNAs as in A.

Europe PMC Funders Author Manuscripts

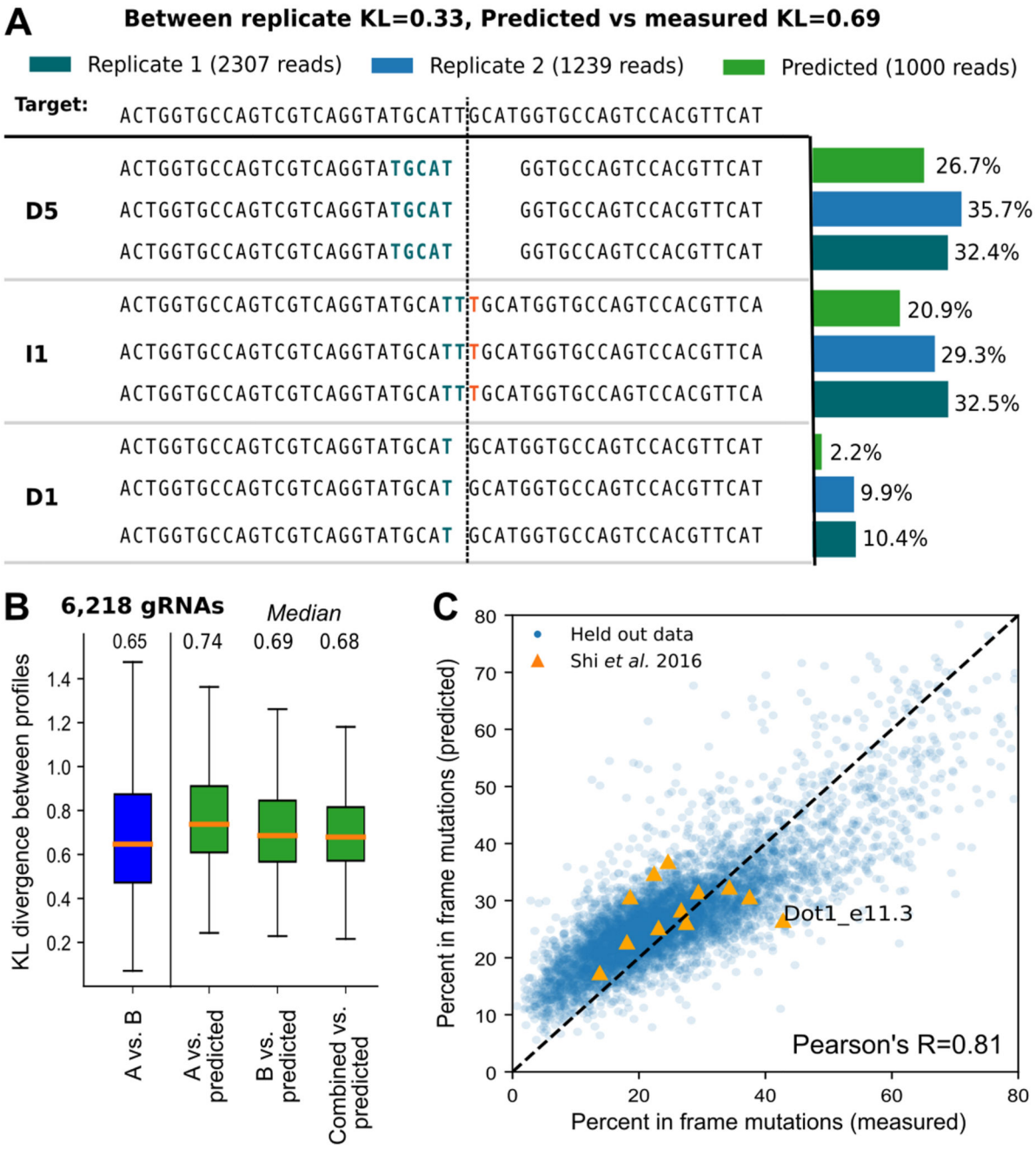

Figure 6. Accurate prediction of repair profiles

A. Example of a repair profile prediction with accuracy close to the test set median (KL=0.69). DNA sequence of the target (top) is edited to produce a range of outcomes in two synthetic replicates (dark green, blue bars) and the corresponding predicted outcomes (green bars). The proportions (x-axis) of the three largest mutational outcomes (“D5” - deletion of size 5 with highlighted size 5 microhomology, “I1” - insertion of a guanine at the cut site, “D1” - deletion of PAM-distal cytosine at the cut site; y-axis) is consistent between

Europe PMC Funders Author Manuscripts

the biological replicates and the prediction. Stretches of microhomology (green) and inserted sequences (red) are highlighted at the cut site (dashed vertical line).

B. Repair profiles can be predicted from sequence alone. Symmetrised Kullback-Leibler divergence (KL, y-axis) between predicted and actual repair profiles (green), as well as between biological replicates A and B (blue; x-axis), with median values denoted above. Box plots: median line with median value marked, quartile box, 95% whiskers. 6,218 gRNAs as in Figure 2C; these were not used in training or hyperparameter selection.

C. Frameshift mutations can be predicted with high accuracy. Measured (x-axis) and predicted (y-axis) percent of mutations that do not produce frameshift mutations for 6,218 held-out gRNAs as in B (blue), and 12 gRNAs that were deep sequenced in (Shi et al. 2015) (orange). Dot1_e11.3 has over 90% deletions of size greater than 30 in the Shi et al sequencing data so we do not expect accurate predictions for this gRNA.