





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

Felicity Allen<sup>1,7</sup>, Luca Crepaldi<sup>1,7</sup>, Clara Alsinet<sup>1</sup>, Alexander J. Strong<sup>1</sup>, Vitalii Kleshchevnikov<sup>1</sup>, Pietro De Angeli<sup>1</sup>, Petra Páleníková<sup>1</sup>, Anton Khodak<sup>1</sup>, Vladimir Kiselev<sup>1</sup>, Michael Kosicki<sup>1</sup>, Andrew R. Bassett<sup>1</sup>, Heather Harding<sup>2</sup>, Yaron Galanty<sup>3,4</sup>, Francisco Muñoz-Martínez<sup>3,4</sup>, Emmanouil Metzakopian<sup>1,5</sup>, Stephen P. Jackson<sup>3,4</sup> & Leopold Parts<sup>1,6</sup>

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, but 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 (gRNAs) 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<sup>9</sup> 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 mutations produced and use these to derive a predictor of Cas9 editing outcomes. Improved understanding of sequence repair will allow better design of gene editing experiments.

CRISPR–Cas9 is a transformative DNA editing technology<sup>1</sup>. It operates by recruiting the Cas9 nuclease to a genomic locus with a protospacer-adjacent motif (PAM) using a short synthetic gRNA with an 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 main repair mechanisms include nonhomologous 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<sup>2,3</sup>. Choice of pathway is influenced by a host of factors, including cell cycle stage and availability of repair enzymes<sup>4,5</sup>. It has been shown that the frequency of alternative Cas9 editing outcomes (the ‘mutational profile’; **Fig. 1**) is largely reproducible and depends on the targeted sequence<sup>6–10</sup>, indicating that the errors in repair occur in a nonrandom 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 *Streptococcus pyogenes*<sup>7</sup>, recently followed up with studies of more target sites<sup>11,12</sup>. More gRNAs (~1,400) were employed in a study that introduced the target and gRNA into cells simultaneously<sup>13</sup>, 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%, yielding insufficient data for a comprehensive analysis. An approach introducing gRNA and target in the same synthetic construct has been used for the Cpf1 nuclease<sup>14</sup> and the *Staphylococcus aureus* Cas9 enzyme<sup>15</sup>. 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 were used to develop an algorithm that predicts indel frequencies, no attempt was made to predict the *S. aureus* Cas9-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 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.

<sup>1</sup>Wellcome Sanger Institute, Hinxton, UK. <sup>2</sup>Cambridge Institute of Medical Research, University of Cambridge, Cambridge, UK. <sup>3</sup>Wellcome/Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge, UK. <sup>4</sup>Department of Biochemistry, University of Cambridge, Cambridge, UK. <sup>5</sup>UK Dementia Research Institute, Cambridge, UK. <sup>6</sup>Department of Computer Science, University of Tartu, Tartu, Estonia. <sup>7</sup>These authors contributed equally to this work. Correspondence should be addressed to F.A. (fa9@sanger.ac.uk) or L.P. (leopold.parts@sanger.ac.uk).

Received 26 July; accepted 12 November; published online 27 November 2018; doi:10.1038/nbt.4317

## RESULTS

### Measuring repair outcomes *en masse*

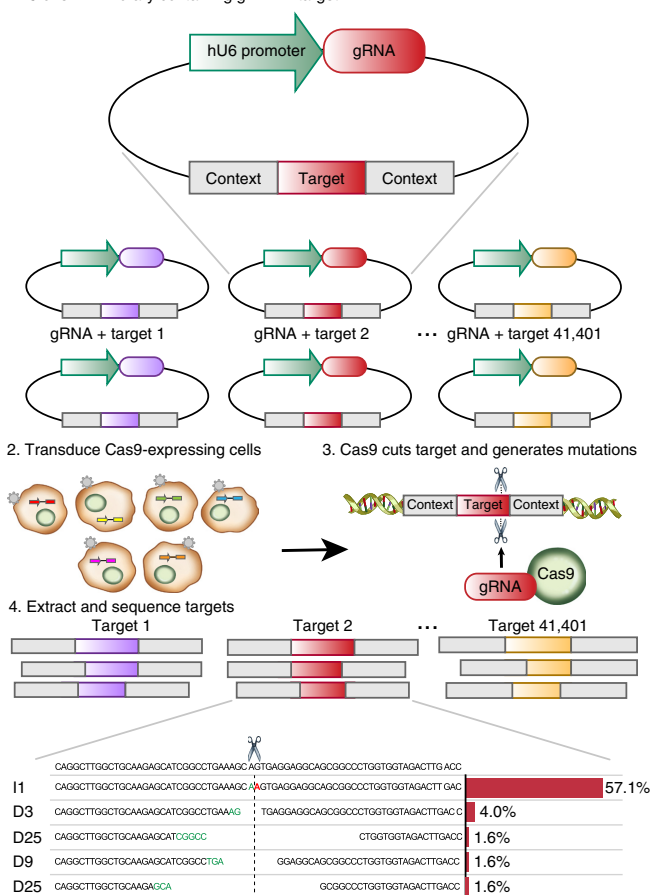
The main hurdle in measuring a large number of repair outcomes is the need to selectively amplify each targeted locus. To circumvent this, we designed a construct that encodes a gRNA expression cassette together with its 23-nt PAM-endowed target sequence within a larger, 79-nt variable context and flanked by common PCR priming sites on both sides (Fig. 1 and Supplementary Fig. 1). 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 (Supplementary Table 1); cultured for 7 d to ensure saturated editing while avoiding drift (Supplementary Figs. 2–4); and then isolated genomic DNA, amplified the target sequence in its context, and sequenced at high coverage (Supplementary Fig. 5) 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 K562 human chronic myelogenous leukemia 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; Fig. 2a–c; 6,218 gRNAs from the conventional set (defined in Online Methods). The fraction of frameshift edits, a factor that is arguably most important for knockout experiments, was also highly correlated between biological replicates (Pearson's  $R = 0.9$ ; Fig. 2d). Together, these results show that the mutational profiles are reproducible and sequence-specific. Given the negligible influence of whether the conventional<sup>16</sup> or improved<sup>17</sup> gRNA scaffolds were used (median KL = 0.77; Fig. 2a,c), the improved version was 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, by van Overbeek *et al.*<sup>7</sup>, in which 223 human genomic targets for 96 unique gRNAs were individually amplified and sequenced ('endogenous outcomes'). The 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 (Fig. 2a), on the whole (median KL = 1.1; Fig. 2b), and for recapitulating frameshift edit fraction (Pearson's  $R = 0.78$ , Fig. 2d). 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 (Fig. 2e). Second, deletions and rearrangements larger than our measurement size limit of 30 nt (Online Methods), which can be prevalent<sup>18</sup>, explain three of the four cases with sufficient read counts that markedly differed (KL > 3). The remaining case (van Overbeek 25) diverges despite high reproducibility between synthetic measurements (Supplementary Fig. 6). Given that our

1. Clone DNA library containing gRNA + target



**Figure 1** Mutational profiles generated by CRISPR–Cas9 and a method for their high-throughput measurement. Constructs containing both a gRNA and its target sequence (matched colors) in variable context (gray 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 is amplified with common primers, and the repair outcomes (location, size and sequence of mutation) determined by deep short-read sequencing (4). I1 indicates insertion of a single base pair, D3 deletion of 3 base pairs, etc. Stretches of microhomology (green) and inserted sequences (red) are highlighted at the cut site (dashed vertical line).

construct only contains 79 nt of local context owing to limitations of oligonucleotide synthesis yet produces very similar outcomes for 94% of measured cases with sufficient reads (67 of 71), this result confirms that sequence surrounding the cut site is the main 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 (Supplementary Table 2).

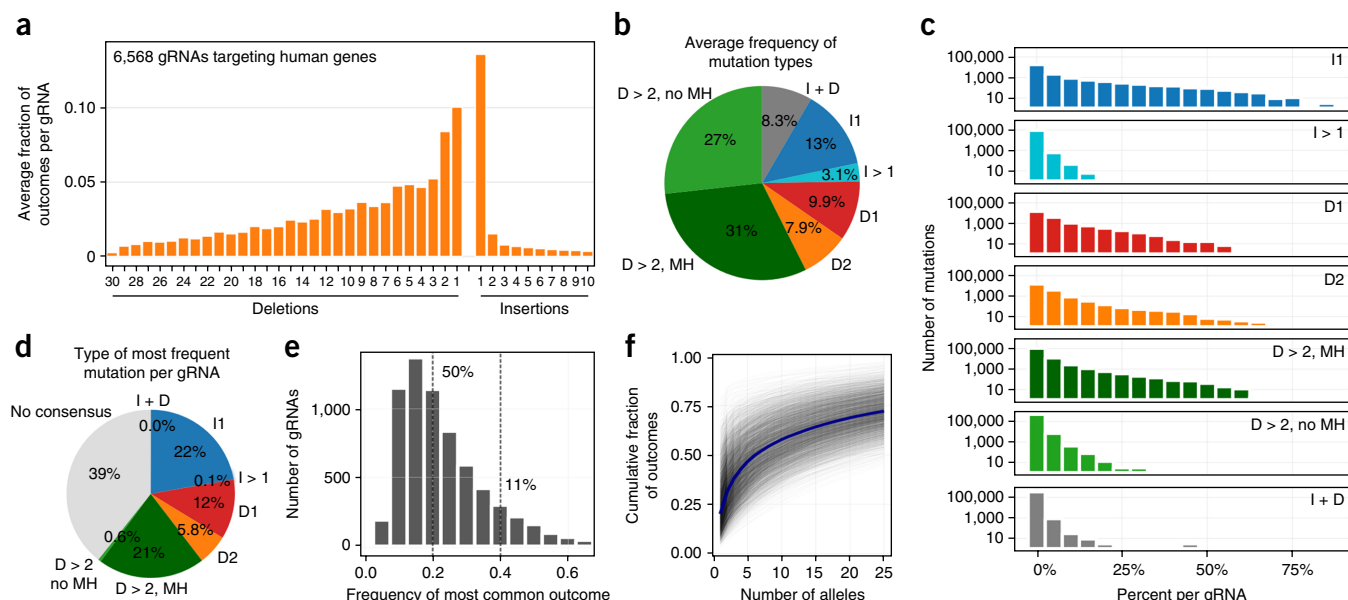
### Repair outcomes in K562 cells are diverse

After concluding that our assay faithfully and reproducibly captures most endogenous mutational outcomes, we surveyed a collection of 6,568 gRNAs that target human genes ('genomic gRNA-targets', Online Methods) and 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 favored over longer ones (**Fig. 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 (**Fig. 3b**). About half of these (31% of the total) occurred between repeating sequences of at least 2 nt ('microhomology'). Deletions of 1 or 2 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).

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 (**Fig. 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.



**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) is averaged across sequence targets present in the genome. **(b)** Editing outcome types are diverse. The percent occurrence per gRNA (area of wedge) of 1-nt 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 insertions plus deletions (I + D, gray), 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<sub>10</sub> 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 percentage 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 (gray lines), and their average (blue line).

Repair outcomes are biased towards particular alleles. The same specific mutation was most frequent in all three biological replicates for over 60% of gRNAs (Fig. 3d), but mutations from different classes were not favored equally. When a consensus existed, it was almost always a single nucleotide insertion (36%), microhomology-mediated deletion of at least 3 nt (34%) or deletion of 1 or 2 bases (30%), and could make up over half of all mutations for that gRNA with reproducible frequency (Pearson’s  $R > 0.83$ ; Fig. 3c and Supplementary Fig. 7). In contrast, whereas deletions of at least 3 nt without microhomology, larger insertions (more than 1 nt) and more complex mutations (insertion plus deletion) are collectively common (38%, Fig. 3b), their frequency for each gRNA is lower and less reproducible (Pearson’s  $R < 0.27$ ; Fig. 3c and Supplementary Fig. 7), so that they form less than 1% of the reproducibly most frequent outcomes (Fig. 3d).

Overall, half of the measured gRNAs had a single outcome that contributed to at least 20% of the observations, and 11% had an outcome that contributed to at least 40% (Fig. 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 (Fig. 3f), indicating a large number of low-frequency events. Some of these may be artifactual, but we expect that to form a minority, as we observed an order of magnitude fewer unique mutations in a control experiment lacking Cas9 (Supplementary Fig. 8). Additionally, frequencies of alleles assayed at different time points (7 and 10 d after infection) from the same replicate are more concordant

than those of biological replicates (Supplementary Fig. 7), indicating resampling 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<sup>6,19</sup>, we first systematically evaluated repair outcomes of targets with different microhomology spans (3–15 nt) and separating distances (0–20 nt). 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 (Fig. 4a). This trend held for all spans of microhomology, but was more pronounced for longer tracts (e.g. Pearson’s  $R = -0.7$  for ten matching bases vs.  $R = -0.2$  for three matching bases; Fig. 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 were generated at a 30% reduced rate if one mismatch was present and 50% reduced rate if two (Fig. 4c). The presence of mutations on 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 (Supplementary Fig. 9) or the rest (Supplementary Fig. 10). We also observed that sequences with low G+C fraction were not as frequently used as repair template as those with higher G+C (Supplementary Fig. 11), 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, as well as, to an unknown extent, in humans<sup>9</sup>. In the following, we consider the alleles that are most frequent for a gRNA in all three replicates ('dominant mutations'; Fig. 3d). Contrary to the lack of directional bias in deletions (Supplementary Figs. 9 and 10), we observed that for over 99% of the 6,572 dominant single base insertions, the PAM-distal nucleotide was repeated (Fig. 4d). Further, for 49% of all gRNAs with a thymine as the PAM-distal base at the cut site, insertion of another thymine was the most frequent outcome, whereas this was the case for only 1.6%, 15% and 28% of gRNAs with a guanine, cytosine and adenine, respectively (Fig. 4e).

A similar strong bias was present for small deletions. We observed that 77% of dominant single base deletions corresponded to removal of a repeating nucleotide at the cut site (Fig. 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 (Fig. 4g).

Repeat removal was also favored for two-base deletions. Half of dominant dinucleotide deletions contracted a repeat of two bases at the cut site, and a further 17% remove a single repeated nucleotide separated by another nucleotide (Fig. 4h). In the remaining cases (31%), both bases were removed from one or other side of the cut, but a single base was never removed from both sides. Notably, if a dinucleotide repeat was present at the cut site of a gRNA, its contraction was very likely to be the dominant repair outcome for that gRNA (up to 77% of gRNAs with an AGAG motif; Fig. 4i), and preference for the PAM-distal pattern was the opposite of a single base deletion, with thymines giving rise to lower rates (10%, 36 of 344), while guanines were preferred (62%, 233 of 377). If alternative sequence configurations were present at the cut site, sequence biases were present (Supplementary Fig. 12), but deleting two bases never dominated for more than 5% of the gRNAs in these other cases (Supplementary Fig. 13).

### Mutational outcomes vary with cell line and some Cas9 modifications

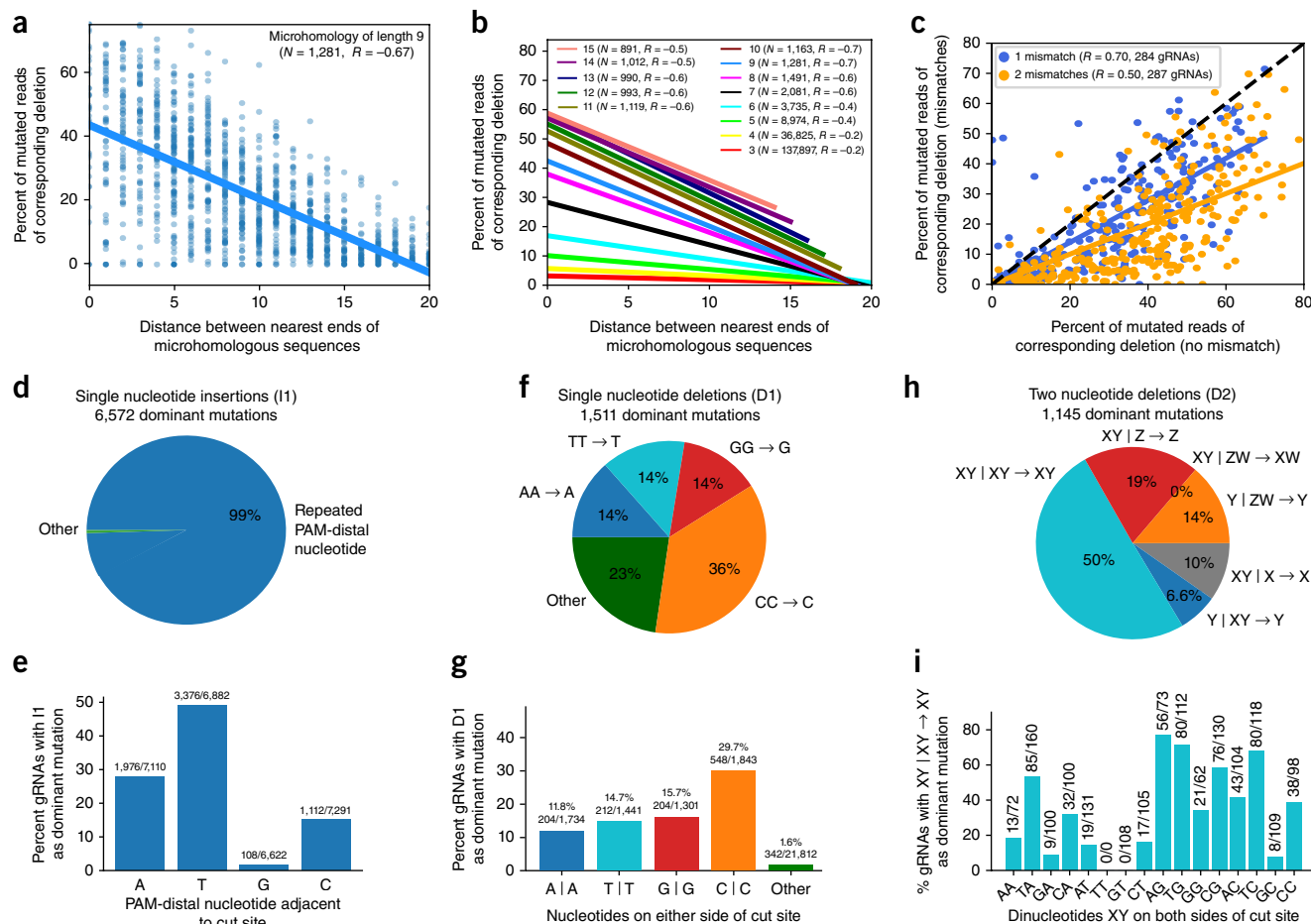
Cells can differ in activity of repair processes and/or DNA sequence, both of which influence double-strand break repair outcomes<sup>20</sup>. We next performed our assay in human induced pluripotent stem cells (iPSCs), mouse embryonic stem cells (mESCs), Chinese hamster ovary (CHO) epithelial cells, 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. Fig. 3), but restricted them to the 3,777 with at least 20 mutated reads in all cell lines (median read numbers in Fig. 5a).

The overall distribution of repair outcome types was not the same across the different cell lines and organisms studied (Fig. 5a), but repair profiles of individual gRNAs remained similar to each other (median KL < 2; Fig. 5b, example in Supplementary Fig. 14). Some relative changes of preferred mutation classes were notable. Large insertions occurred more frequently in stem cells, with 2.6-fold and 1.7-fold increases in human iPSCs and mESCs, respectively, over K562 levels. However, the frequency of these mutations remained low and non-reproducible (Supplementary Fig. 15), which also explains the increase in overall between-replicate divergence (Fig. 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 was present in CHO cells, whereas both iPSCs and mESCs favored microhomology-mediated deletions at the expense of single base insertions (Fig. 5a). This trend was recapitulated in the mutation class of the dominant mutation for each gRNA, which changed depending on the cell line (Fig. 5d). We found little influence of the genetic background on the link between microhomology and repair outcomes (Fig. 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)<sup>21</sup> and Cas9 fused to three-prime repair exonuclease 2 (TREX2), which is known to increase deletion size<sup>13,22</sup>. eCas9 behaved similarly to Cas9 (Fig. 5a,b), albeit with a slower editing saturation (Supplementary Fig. 2), whereas outcomes in the Cas9-TREX2 fusion protein line were markedly different from the others (Fig. 5b). Cas9-TREX2 mutations were shifted towards larger deletions (Fig. 5d,e) and favored ligation of the intact PAM-proximal side with a deletion on the PAM-distal side (Supplementary Fig. 16), at the expense of frequent microhomology-mediated deletions (Fig. 5c and Supplementary Fig. 17). The generation of less biased and larger deletions, as produced 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 the Cas9-TREX2 fusion, but did increase the frequency of single base deletions while reducing microhomology-mediated ones (Supplementary Figs. 17 and 18). Combined, these data are consistent with the function of TREX2 as an exonuclease<sup>23</sup> that promotes repair via the canonical nonhomologous end joining pathway, resulting in small deletions that are not mediated by microhomology<sup>24,25</sup>.

### Repair outcomes can be accurately predicted

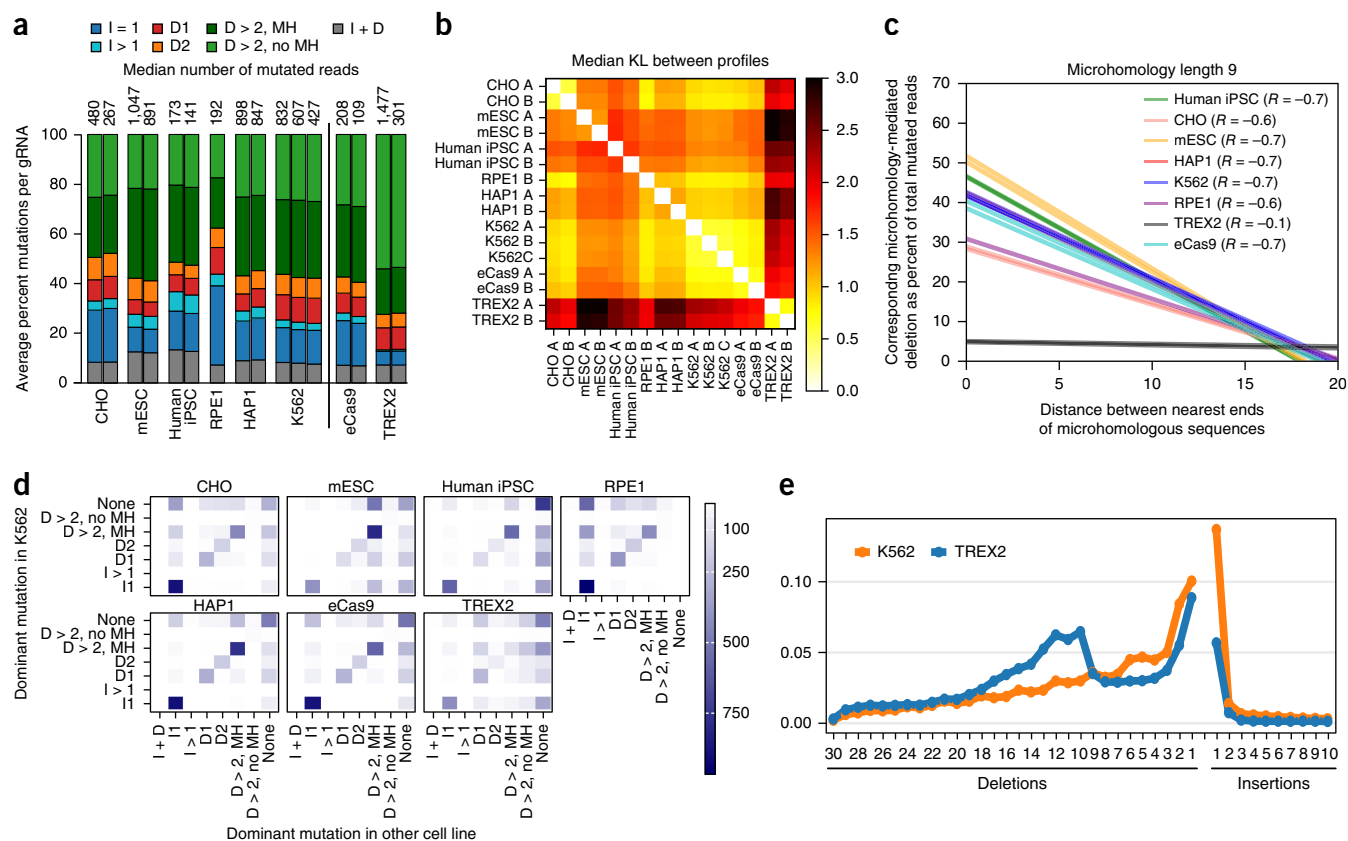
So far, we have demonstrated that the repair outcomes are reproducible, biased, dependent on the local sequence and mostly consistent across genetic backgrounds. These observations suggest that mutations generated by Cas9 ought to be predictable from sequence alone. To test this hypothesis, we developed a computational predictor of the mutational outcomes of a given gRNA, which we call FORECasT (favored 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 (Fig. 6a).



**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 figure. **(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 figure. **(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). **(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 adenine (blue), repeating thymine (teal), repeating guanine (red), repeating cytosine (orange) or a base from a nonrepeat (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 favor 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 X, gray), or delete a PAM-proximal nucleotide flanked by a repeating base (Y | XY Y, blue). X, Y, Z and W represent any nucleotide; the vertical bar represents the 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).

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; **Fig. 6b**; examples of predictions at the quartiles and whiskers in **Supplementary Fig. 19**). Frequencies of individual mutations that were reproducible (1-nt insertions and 1-, 2- and >2-nt deletions

with microhomology) were also correspondingly well predicted (**Supplementary Fig. 20**). As a consequence, we could accurately predict the percentage of mutations that do not disrupt the reading frame on held-out validation data (Pearson's  $R = 0.81$  for prediction vs. 0.89 for replicates; **Figs. 2d** and **6c**). 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 mESCs; **Supplementary**



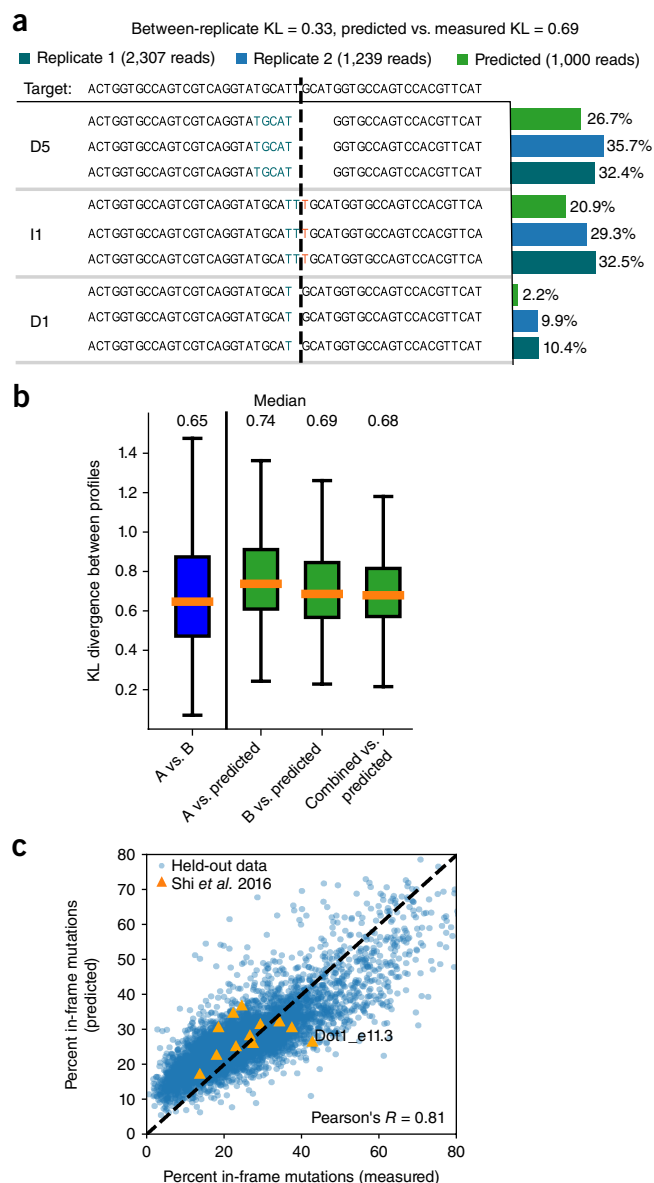
**Figure 5** Differences between editing outcomes in K562-Cas9 and other cell lines and effector proteins. **(a)** Genetic background influences editing outcomes. Average per-gRNA frequency of different types of editing outcomes in 3,777 gRNAs (y axis; colors as **3b**) for CHO cells, mESCs, human iPSCs, human RPE-1 cells, human HAP1 cells, K562 cells and K562 cells with alternative Cas9 proteins: 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 KL divergence between repair profiles (black to white color range, as in **Fig. 2b**) in different tested lines (x and y axes). 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 given in the figure. gRNAs as in **Figure 4b**, restricted to those 822 gRNAs with microhomology 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 toward microhomology-mediated deletions in stem cells and toward 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 all replicates. gRNAs as in **a**. RPE data are based on one replicate, K562 on three, all other cell lines on two. **(e)** Cas9-TREX2 fusion protein favors larger deletions than 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**.

**Fig. 21**), with the magnitude of discrepancies consistent with shifts in the repair profiles (**Fig. 5a,b**).

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 thymidine nucleotide (**Supplementary Table 3**). Individual deletion-related features (over 2,000 total) had lower weights, most likely due to their larger quantity. Substantial biases in feature weights highlight the expected microhomology-related properties explored above, among 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 knockout experiments and large-scale screens. We predicted mutational outcomes for gRNAs targeting ten genes<sup>26</sup>, but while the

estimated fraction of frameshift edits is concordant with the measurements where available (12 gRNAs; **Fig. 6c**, **Supplementary Fig. 22** and **Supplementary Table 4**), it does not explain the observed phenotypic variation (**Supplementary Fig. 23**). We also calculated correlations between the predicted fraction of out-of-frame mutations and gRNA efficacy for gRNAs targeting essential genes in three large-scale screening datasets (Online Methods) and found a significant, albeit small, link (**Supplementary Fig. 24**). 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 ref. 26, we further observed a weak but consistent association between the predicted fraction of out-of-frame mutations and phenotypic effect for gRNAs that target outside protein domains (**Supplementary Fig. 25**), indicating that it is more important to disrupt the reading frame in those regions.



**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 (teal and blue bars) and the corresponding predicted outcomes (green bars). The proportions (x axis) of the three largest mutational outcomes (where D5 indicates a deletion of size 5 with highlighted size 5 microhomology, I1 an insertion of a guanine at the cut site, D1 a deletion of the PAM-distal cytosine at the cut site; y axis) is consistent between 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. Symmetrized KL divergence (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 by Shi et al.<sup>26</sup> (orange). Dot1\_e11.3 has over 90% deletions of size greater than 30 in the Shi et al.<sup>26</sup> sequencing data, so we do not expect accurate predictions for this gRNA.

These results demonstrate that repair outcomes can be predicted from sequence alone and in a manner that is expected to generalize to all sites in the genome. We have made the predictor available as a web tool 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, to our knowledge, the 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-directed repair, which shares the initial resection step<sup>4</sup>, whereas favoring of single base insertions in CHO and RPE-1 lines indicates elevated canonical end-joining activity. The higher incidence of large insertions in stem cells could similarly be explained by aberrant homology-directed repair, wherein strand invasion occurs in the wrong place, such that DNA synthesis before strand displacement leads to additional sequence being inserted.

The strong sequence biases observed here for single nucleotide insertions, as previously also seen in yeast and to some extent humans<sup>9</sup>, could be explained by a model whereby the Cas9 protein stays bound to the PAM-proximal side of the cut while the staggered 1-nt overhang on the PAM-distal side<sup>10,27</sup> is filled by DNA polymerase and re-ligated via the nonhomologous end joining pathway<sup>9,28</sup>. 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<sup>18</sup>, 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, such as Huntington's disease or fragile X syndrome, are due to expansions of short tandem repeats<sup>29</sup>. 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<sup>30</sup>. In the future, contraction of such rogue expansions could be explored as a therapy option, as the low-efficacy allele replacement is not required; simply generating a double-strand break would shorten the pathogenic repeat. Indeed, a few preliminary efforts in this direction have already given promising results<sup>31–33</sup>, but given the possible unintentional genomic damage<sup>18</sup>, 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 custom edits.



## METHODS

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

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

## ACKNOWLEDGMENTS

We 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. F.A. was supported by a Royal Commission for the Exhibition of 1851 Research Fellowship. L.P. was supported by Wellcome (206194) and the Estonian Research Council (IUT 34-4). H.P.H. was supported by a Wellcome Trust grant (200848/Z/16/Z) and a Wellcome Trust Strategic Award to the Cambridge Institute for Medical Research (100140). Y.G. is funded by Cancer Research UK C6/A18796 and Wellcome Trust Investigator Award 206388/Z/17/Z in the Jackson laboratory. F.M.M. was funded by a Marie Curie Intra-European Fellowship, project number 626375, DDR SYNIVA, and by Wellcome Trust Investigator Award 206388/Z/17/Z and an AstraZeneca Collaborative Award in the Jackson laboratory.

## AUTHOR CONTRIBUTIONS

F.A.: designed experiments, analyzed data, wrote paper. L.C.: designed experiments, performed experiments, wrote paper. C.A.: performed experiments in human iPSCs. A.J.S., E.M.: performed experiments in mouse ESCs. V. Kleshchevnikov: analyzed data, wrote paper. A.K., V. Kiselev: created web server. P.D.A., P.P.: performed experiments. M.K., A.R.B.: generated TREX2 constructs. H.H.: generated CHO-Cas9 line. Y.G., F.M.-M., S.P.J.: generated RPE-1-Cas9 and HAP1-Cas9 lines. L.P.: designed experiments, contributed to data analysis, wrote paper. All authors contributed to drafting the manuscript.

## COMPETING INTERESTS

The authors declare no competing interests.

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

- Doudna, J.A. & Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
- Chiruvella, K.K., Liang, Z. & Wilson, T.E. Repair of double-strand breaks by end joining. *Cold Spring Harb. Perspect. Biol.* **5**, a012757 (2013).
- Her, J. & Bunting, S.F. How cells ensure correct repair of DNA double-strand breaks. *J. Biol. Chem.* **293**, 10502–10511 (2018).
- Truong, L.N. *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. USA* **110**, 7720–7725 (2013).
- Shibata, A. Regulation of repair pathway choice at two-ended DNA double-strand breaks. *Mutat. Res.* **803–805**, 51–55 (2017).
- Bae, S., Kweon, J., Kim, H.S. & Kim, J.-S. Microhomology-based choice of Cas9 nuclease target sites. *Nat. Methods* **11**, 705–706 (2014).
- van Overbeek, M. *et al.* DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks. *Mol. Cell* **63**, 633–646 (2016).
- 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**, 267–273 (2014).
- Lemos, B.R. *et al.* CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. *Proc. Natl. Acad. Sci. USA* **115**, E2040–E2047 (2018).
- 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).
- Taheri-Ghahfarokhi, A. *et al.* Decoding non-random mutational signatures at Cas9 targeted sites. *Nucleic Acids Res.* **46**, 8417–8434 (2018).
- Chakrabarti, A.M. *et al.* Target-specific precision of CRISPR-mediated genome editing. Preprint at *bioRxiv* <https://doi.org/10.1101/387027> (2018).
- Chari, R., Mali, P., Moosburner, M. & Church, G.M. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods* **12**, 823–826 (2015).
- Kim, H.K. *et al.* In vivo high-throughput profiling of CRISPR-Cpf1 activity. *Nat. Methods* **14**, 153–159 (2017).
- Tycko, J. *et al.* Pairwise library screen systematically interrogates *Staphylococcus aureus* Cas9 specificity in human cells. *Nat. Commun.* **9**, 2962 (2018).
- Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
- Chen, B. *et al.* Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).
- 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**, 765–771 (2018).
- Cho, S.W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24**, 132–141 (2014).
- Gallagher, D.N. & Haber, J.E. Repair of a site-specific DNA cleavage: old-school lessons for Cas9-mediated gene editing. *ACS Chem. Biol.* **13**, 397–405 (2018).
- Slaymaker, I.M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).
- Bothmer, A. *et al.* Characterization of the interplay between DNA repair and CRISPR/Cas9-induced DNA lesions at an endogenous locus. *Nat. Commun.* **8**, 13905 (2017).
- Mazur, D.J. & Perrino, F.W. Excision of 3' termini by the Trex1 and TREX2 3'5' exonucleases. Characterization of the recombinant proteins. *J. Biol. Chem.* **276**, 17022–17029 (2001).
- Bhargava, R., Carson, C.R., Lee, G. & Stark, J.M. Contribution of canonical nonhomologous end joining to chromosomal rearrangements is enhanced by ATM kinase deficiency. *Proc. Natl. Acad. Sci. USA* **114**, 728–733 (2017).
- Certo, M.T. *et al.* Coupling endonucleases with DNA end-processing enzymes to drive gene disruption. *Nat. Methods* **9**, 973–975 (2012).
- Shi, J. *et al.* Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. *Nat. Biotechnol.* **33**, 661–667 (2015).
- Zuo, Z. & Liu, J. Cas9-catalyzed DNA cleavage generates staggered ends: evidence from molecular dynamics simulations. *Sci. Rep.* **5**, 37584 (2016).
- Richardson, C.D., Ray, G.J., DeWitt, M.A., Curie, G.L. & Corn, J.E. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat. Biotechnol.* **34**, 339–344 (2016).
- Sutherland, G.R. & Richards, R.I. Simple tandem DNA repeats and human genetic disease. *Proc. Natl. Acad. Sci. USA* **92**, 3636–3641 (1995).
- Gu, Y., Shen, Y., Gibbs, R.A. & Nelson, D.L. Identification of FMR2, a novel gene associated with the FRAXE CCG repeat and CpG island. *Nat. Genet.* **13**, 109–113 (1996).
- Cinesi, C., Aeschbach, L., Yang, B. & Dion, V. Contracting CAG/CTG repeats using the CRISPR-Cas9 nickase. *Nat. Commun.* **7**, 13272 (2016).
- Mahadevan, M.S. *et al.* Reversible model of RNA toxicity and cardiac conduction defects in myotonic dystrophy. *Nat. Genet.* **38**, 1066–1070 (2006).
- Park, C.-Y. *et al.* Reversion of FMR1 methylation and silencing by editing the triplet repeats in fragile X iPSC-derived neurons. *Cell Rep.* **13**, 234–241 (2015).

## ONLINE METHODS

**Selection of guides and targets.** We compiled our library of 41,630 total gRNA–target pairs (**Supplementary Data 1**) from five sublibraries aimed at testing different aspects of repair outcome generation:

**Endogenous gRNA–target pairs.** We included 86 gRNAs used in ref. 7 ('endogenous targets') that were compatible with our library cloning method (see below), of which 9 were lost during cloning, giving 77 in total.

**Genomic gRNA–target pairs** 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<sup>34</sup>. Of these, 5,194 were selected because they were also present in the library used in refs. 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. A further 903 guide–target pairs were included from within the set used in ref. 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 nonessential genes, or were in the Endogenous set above.

**Explorative gRNA–target pairs** 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) and 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 **Supplementary Data 1**, and an overview is given in **Supplementary Table 5**.

**Microhomology mismatch gRNA–target pairs** 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.

**Conventional scaffold gRNA–target pairs** All of the above subsets used the improved gRNA scaffold<sup>17</sup>. 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), 2,364 Explorative), were ordered as separate oligonucleotides in the purchased pool and were independently cloned with the alternative conventional gRNA scaffold<sup>16</sup>. These gRNAs allowed assessment of the impact of a difference in gRNA scaffold (which appears very small; **Fig. 2c**) as well as providing independently synthesized and constructed repeat measurements of the same gRNAs.

Every target is uniquely barcoded by a 10-nt sequence at both the 3' and 5' ends (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 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 guanine in the gRNA (but not the target) in the position 20 nt before the PAM, for improved expression of the gRNA from the hU6 promoter.

**Construction of the 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<sup>34</sup> (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, **Supplementary Table 6**) 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 harboring two BbsI restriction sites (**Supplementary Fig. 1**), enclosed by priming sites, and using all remaining 79 nt to randomize the sequence context of the target. Primer pairs P5-P6 and P7-P8 (**Supplementary Table 6**) were used to amplify oligonucleotides compatible with the conventional or improved scaffold, respectively. Gibson assembly<sup>36</sup> 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. A 1:1 molar ratio was mixed in three reactions, incubated 1 h at 50 °C, 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 linearization with BbsI at 37 °C for 2 h. The resulting 296-bp linear fragment was ligated into scaffoldless pKLV2-U6(BbsI)-PKGpuro2ABFP-W. Ligations (T4 DNA ligase, NEB) were performed in triplicate, pooled and used in up to ten electroporation reactions to maximize 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 sequence<sup>37</sup> with a GGS linker or an intervening T2A peptide. These were cloned by Gibson assembly into a piggyBac vector (pKLV-Cas9) driven by an EFS promoter and containing a blasticidin-selectable marker. GenBank files of the final vectors are provided in **Supplementary Data 2**.

**Cell culture.** K562, K562-Cas9 (a kind gift by E. 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 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 medium was changed daily throughout expansion and all experiments. All cell lines were cultured at 37 °C, 5% CO<sub>2</sub>.

**Lentiviral 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 of 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 the presence of 8 µg/ml Polybrene (hexadimethrine bromide, Sigma) followed by centrifugation for 30 min at 1,000g. E14TG2a mouse ESCs transduction was performed incubating cells in suspension for 30 min in presence of 8 µg/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 8 µg/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) (Addgene 71814)<sup>21</sup>, Cas9-TREX2 and Cas9-2A-TREX2 vectors to generate K562-eCas9, K562-Cas9-TREX2 and K562-Cas9-2A-TREX2, respectively. E14TG2a-Cas9 cells were generated by transducing E14TG2a cells with pKLV2-EF1aBsd2ACas9-W (Addgene 67978)<sup>34</sup>.

**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 500× to 1,600×. Total number of cells, MOI and coverage for each screen are listed in **Supplementary Table 1**. 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 d after infection, with a small number of samples further maintained for up to 20 d to evaluate the effect of time-point choice (**Supplementary Table 1**). 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 −20 °C before 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 55 °C for 16 h. DNA was extracted by adding one volume of isopropanol followed by spooling, washed twice in 70% ethanol, centrifuged and resuspended in 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 (**Supplementary Table 6**) 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 taking 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, and 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 (ref. 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 confirmed that the read could be aligned to the template in such a way that at least 80% of the read characters have a 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 oligonucleotides, we checked each sample to ensure that the per-oligonucleotide

log<sub>2</sub>(read count) values (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 (Pearson’s  $R > 0.95$ , computed using `scipy.stats.pearsonr`<sup>39</sup>) 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 oligonucleotide 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 sublibraries 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. 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<sup>40</sup>) and may therefore be expected to decrease in coverage as the assay goes on, the limited duration of the assay (7 d) 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 that target essential genes (as defined by ref. 40) compared to those that do not (**Supplementary Fig. 26**).

**Repair profile comparisons.** We store the repair profile as a collection of read counts per indel, where each indel is characterized 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 KL divergence is calculated as

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

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 **Supplementary Figure 19a–d**.

For frameshift 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 **Supplementary Figures 2 and 5**.

**Repair profile analysis.** We classified all indels into types I1 (size 1 insertion), I > 1 (larger insertions), D1 (size 1 deletion), D2 (size 2 deletion),



I + D (insertion and deletion) and D > 2 (larger deletions). For deletions of size greater than 2 we assigned a causal mechanism of the event as likely generated by microhomology if at least 2 nt of matching sequence were present on either side of the cut and likely not generated by microhomology otherwise. To analyze indel prevalence by size (Fig. 3a), we classified indels into deletions of 1–30 nt and insertions of 1–10 nt. In this case, information about deletions larger than 30 and insertions larger than 10 nt was excluded from the analysis because they are not well detected by our method. In this and other results presenting accumulated measurements across gRNAs (Figs. 3a–c and 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 site with a left edge from position 1 to the 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 modeled the probability of each possible outcome using a logistic that ensures the sum of all possible outcomes for a given gRNA sums to 1. That is, the probability of the  $j$ th mutation for a given gRNA is modeled as

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

where  $\theta$  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, nonsymmetric KL divergence of these probabilities when compared to the measured proportions, by computing closed-form partial gradients with respect to the  $k$ th parameter  $\theta_k$  and using L-BFGS-B within `scipy.optimize.minimize` to perform gradient descent optimization of this metric<sup>39,41</sup>.

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  $N = 50, 100, 200, 500, 1,000$  and  $5,000$  gRNAs from this set and assigning these to training and test sets, respectively, and repeating this three times for each hyperparameter and training set size. With  $5,000$  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 (Supplementary Fig. 27). We used the Conventional scaffold gRNA–targets set as a held-out validation set, predicting profiles by applying the associated predicted probabilities to generate  $1,000$  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 Figures 2d and 6 summed counts from both  $800 \times \text{DPI7}$  replicates from the K562 Improved scaffold samples, whereas replicate B used the single replicate  $1600 \times \text{DPI7}$  sample.

**Endogenous data processing.** We collected the raw read data from the SRA archives referenced by refs. 7 and 26 and reprocessed the generated mutations for each gRNA using the same custom alignment program we used for our own data. For the van Overbeek *et al.*<sup>7</sup> data, we used the data they collected in K562 cells 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 data for the heights and locations of the Shi *et al.*<sup>26</sup> data used in Supplementary Figure 23 were obtained via personal communication with the authors; we thank them for their assistance.

To analyze the influence of chromatin on the results in Figure 2e, we downloaded the ChromHMM<sup>42</sup> chromatin state assignments for K562 cell line as measured in ENCODE<sup>43</sup> from <http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeAwgSegmentation/wgEncodeAwgSegmentation->

[ChromHMMK562.bed.gz](http://ChromHMMK562.bed.gz) and used `bedtools`<sup>44</sup> to overlap them with the target locations in van Overbeek *et al.*<sup>7</sup>.

**Frameshift 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)<sup>45</sup>, Tzelepis *et al.* 2016 (Yusa v1.0 library)<sup>34</sup> and Aguirre *et al.* 2016 (GeCKO v2 library)<sup>46</sup>. Our metric for relative gRNA efficacy was the gRNA scores inferred by JACKS<sup>47</sup>, 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<sup>40</sup>. Pearson’s  $R$  coefficients were calculated using `scipy.stats.pearsonr`<sup>39</sup>.

To assess the importance of frameshift 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 the R package `ensembl` and annotation package `EnsDb.Hsapiens.v75` based on ENSEMBL version 75 and GRCh37 genome assembly<sup>48</sup>.

**Code availability.** The newly developed code for all analyses has been peer reviewed and is available at <https://github.com/felicityallen/SelfTarget> and as **Supplementary Software**. An executable version is also provided on Code Ocean at <https://codeocean.com/2018/11/14/predicting-the-mutations-generated-by-repair-of-cas9-induced-dsbs/metadata> (ref. 49). The code is distributed under the MIT license <https://opensource.org/licenses/MIT>.

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

**Data availability.** Raw sequence data are available at European Nucleotide Archive (Project [PRJEB29746](https://www.ebi.ac.uk/ena/record/PRJEB29746) sample accessions provided in **Supplementary Data 3**). Processed mutational profiles are provided on [https://figshare.com/articles/processed\\_mutational\\_profiles/7312067](https://figshare.com/articles/processed_mutational_profiles/7312067) (<https://doi.org/10.6084/m9.figshare.7312067>).

34. Tzelepis, K. *et al.* A CRISPR dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid leukemia. *Cell Rep.* **17**, 1193–1205 (2016).
35. Wang, T., Wei, J.J., Sabatini, D.M. & Lander, E.S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
36. Gibson, D.G. Enzymatic assembly of overlapping DNA fragments. *Methods Enzymol.* **498**, 349–361 (2011).
37. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
38. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**, 614–620 (2014).
39. Jones, E., Oliphant, T. & Peterson, P. SciPy: open source scientific tools for Python. *SciPy* <http://www.scipy.org> (2001, accessed 10 January 2018).
40. Hart, T. *et al.* Evaluation and design of genome-wide CRISPR/SpCas9 knockout screens. *G3 (Bethesda)* **7**, 2719–2727 (2017).
41. Zhu, C., Byrd, R.H., Lu, P. & Nocedal, J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.* **23**, 550–560 (1997).
42. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
43. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
44. Quinlan, A.R. BEDTools: The Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**, 11.12.1–11.12.34 (2014).
45. Meyers, R.M. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).
46. Aguirre, A.J. *et al.* Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer Discov.* **6**, 914–929 (2016).
47. Allen, F. *et al.* JACKS: joint analysis of CRISPR/Cas9 knock-out screens. Preprint at *bioRxiv* <https://doi.org/10.1101/285114> (2018).
48. Zerbino, D.R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
49. Allen, F. *et al.* Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Code Ocean* <https://doi.org/10.24433/CO.6bc7bcae-d736-475b-bae5-00ca0562d401> (2018).



## Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

### Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

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

Our web collection on [statistics for biologists](#) may be useful.

### Software and code

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

Data collection

The code and dependencies used during data collection and analysis are available at <https://github.com/felicityallen/selftarget> and in a code ocean capsule at <https://codeocean.com/capsule/6bc7bcae-d736-475b-bae5-00ca0562d401>

Data analysis

The code and dependencies used during data collection and analysis are available at <https://github.com/felicityallen/selftarget> and in a code ocean capsule at <https://codeocean.com/capsule/6bc7bcae-d736-475b-bae5-00ca0562d401>. We used R v3.5, Python v3.6 and C++. R ensembl package used EnsDb.Hsapiens.v75

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

### Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data is provided in Supplementary Tables with the manuscript, and the raw sequences in the European Nucleotide Archive (PRJEB12405 / ERP013879).

## Field-specific reporting

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

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

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

## Life sciences study design

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

Sample size	Sample sizes were picked to have adequate signal to noise ratio and repeatability in the measurements.
Data exclusions	The data exclusion criteria are provided in the Methods section of the manuscript. Briefly, we excluded low quality samples that did not resemble the control as expected, as well as excluded individual data points that had insufficient quality (low number of sequencing reads).
Replication	A full list of replicates is given in Supplementary tables. All measurements were performed in duplicate or triplicate, except RPE-1 samples, for which two replicates were performed, but only one passed quality control.
Randomization	Samples were not randomized.
Blinding	Investigators were not blinded.

## Reporting for specific materials, systems and methods

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Unique biological materials

Policy information about [availability of materials](#)

Obtaining unique materials	Describe any restrictions on the availability of unique materials OR confirm that all unique materials used are readily available from the authors or from standard commercial sources (and specify these sources).
----------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

## Antibodies

Antibodies used	Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.
Validation	Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

## Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	Cas9-expressing K562, 293FT, CHO, HAP1, RPE-1, and E14TG2a cell lines derive from commercially available lines and have been generated using lentiviral vectors for the stable expression of SpCas9. iPSCs were originally derived by Ludovic Vallier and colleagues at the Laboratory for Regenerative Medicine, University of Cambridge, Cambridge, United Kingdom (Tamir Rashid et al., J Clin Invest. 2010;120(9):3127-3136).
---------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Authentication	None of the cell lines used were authenticated in our lab
Mycoplasma contamination	All cell lines have been tested for mycoplasma contamination and resulted negative.
Commonly misidentified lines (See <a href="#">ICLAC</a> register)	None of the cells used in this study are listed in the ICLAC register.

## Palaeontology

Specimen provenance	Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).
Specimen deposition	Indicate where the specimens have been deposited to permit free access by other researchers.
Dating methods	If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.
Wild animals	Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.
Field-collected samples	For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

## Human research participants

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

Population characteristics	Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."
Recruitment	Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

## ChIP-seq

### Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links May remain private before publication.	For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.
Files in database submission	Provide a list of all files available in the database submission.
Genome browser session (e.g. <a href="#">UCSC</a> )	Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

### Methodology

Replicates	Describe the experimental replicates, specifying number, type and replicate agreement.
Sequencing depth	Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.
Antibodies	Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

## Peak calling parameters

*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*

## Data quality

*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*

## Software

*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

## Flow Cytometry

### Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

### Methodology

## Sample preparation

*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*

## Instrument

*Identify the instrument used for data collection, specifying make and model number.*

## Software

*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*

## Cell population abundance

*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*

## Gating strategy

*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

- ☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

## Magnetic resonance imaging

### Experimental design

## Design type

*Indicate task or resting state; event-related or block design.*

## Design specifications

*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*

## Behavioral performance measures

*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

### Acquisition

## Imaging type(s)

*Specify: functional, structural, diffusion, perfusion.*

## Field strength

*Specify in Tesla*

## Sequence &amp; imaging parameters

*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

## Area of acquisition

*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

## Diffusion MRI

☐ Used

☐ Not used

### Preprocessing

## Preprocessing software

*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

## Normalization

*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*



## Normalization template

*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

## Noise and artifact removal

*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

## Volume censoring

*Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

## Statistical modeling &amp; inference

## Model type and settings

*Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

## Effect(s) tested

*Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference  
(See [Eklund et al. 2016](#))

*Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

## Correction

*Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

## Models &amp; analysis

n/a | Involved in the study

- ☐ ☐ Functional and/or effective connectivity
- ☐ ☐ Graph analysis
- ☐ ☐ Multivariate modeling or predictive analysis

## Functional and/or effective connectivity

*Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*

## Graph analysis

*Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*

## Multivariate modeling and predictive analysis

*Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*

## Editorial Policy Checklist

This form is used to ensure compliance with Nature Research editorial policies related to research ethics and reproducibility. For further information, please see our [Authors & Referees](#) site. All relevant questions on the form must be answered.

### ► Competing interests

Policy information about [competing interests](#)

#### Competing interests declaration

In the interest of transparency and to help readers form their own judgements of potential bias, Nature Research journals require authors to declare any competing financial and/or non-financial interest in relation to the work described in the submitted manuscript.

☒ No, I declare that the authors have no competing financial or non-financial interests as defined by Nature Research.

☐ Yes, I declare that the authors have a competing interest as defined by Nature Research

### ► Data availability

Policy information about [availability of data](#)

#### Data availability statement

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

☒ A full data availability statement is included in the manuscript.

#### Mandated accession codes ([where applicable](#))

Confirm that all relevant data are deposited into a public repository and that accession codes are provided.

☒ All relevant accession codes are provided ☐ Accession codes will be available before publication ☐ No data with mandated deposition

### ► Data presentation

#### Image integrity

☒ Confirm that all images comply with our [image integrity policy](#).

Unprocessed data must be provided upon request. Please double-check figure assembly to ensure that all panels are accurate (e.g. all labels are correct, no inadvertent duplications have occurred during preparation, etc.).

#### Data distribution

Present data in a format that shows data distribution (dot-plots or box-and-whisker plots).

Define all box-plot elements (e.g. center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers).

If using bar graphs, overlay the corresponding dot plots.

☒ Confirm that all data presentation meets these requirements and that individual data points are shown.

## Specific policy considerations

Some types of research require additional policy disclosures. Please indicate whether these apply to your study. If you are not certain, please read the appropriate section before selecting a response.

Does not apply	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Custom software or computer code
<input checked="" type="checkbox"/>	<input type="checkbox"/> Macromolecular structural data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Research animals and/or animal-derived materials that require ethical approval
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

## ► Code availability

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

### Code availability statement

For all studies using custom code, the Methods section must include a statement under the heading "Code availability" describing how readers can access the code, including any access restrictions.

☒ A full code availability statement is included in the manuscript

## ► Macromolecular structural data

Policy information about [special considerations](#) for specific types of data

### Validation report

☐ For all macromolecular structures studied, confirm that you have provided an official validation report from [wwPDB](#).

## ► Research animals

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

### Ethical compliance

☐ Confirm that you have complied with all relevant ethical regulations and that a statement affirming this is included in the manuscript.

### Ethics committee

☐ Confirm that the manuscript states the name(s) of the board and institution that approved the study protocol.

## ► Human research participants

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

### Ethical compliance

☐ Confirm that you have complied with all relevant ethical regulations and that a statement affirming this is included in the manuscript.

### Ethics committee

Confirm that the manuscript states the name(s) of the board and/or institution that:

☐ Approved the study protocol      -OR-      ☐ Provided guidelines for study procedures (if protocol approval is not required)

### Informed consent

☐ Confirm that informed consent was obtained from all participants.

### Identifiable images

For publication of identifiable images of research participants, confirm that consent to publish was obtained and is noted in the Methods.

Authors must ensure that consent meets the conditions set out in the [Nature Research participant release form](#).

☐ Yes      ☐ No identifiable images of human research participants

## ► Clinical studies

Policy information about [clinical studies](#)

### Clinical trial registration

☐ Confirm that you have provided the trial registration number from [ClinicalTrials.gov](#) or an equivalent agency in the manuscript.

### Phase 2 and 3 randomized controlled trials

Confirm that you have provided the [CONSORT checklist](#) with your submission.

☐ Yes      ☐ No      ☐ Not a phase 2/3 randomized controlled trial

### Tumor marker prognostic studies

Did you follow the [REMARK reporting guidelines](#)?

☐ Yes      ☐ No      ☐ Not a tumor marker prognostic study

I certify that all the above information is complete and correct.

Typed signature Leopold Parts, Felicity Allen Date Sep 24, 2018

