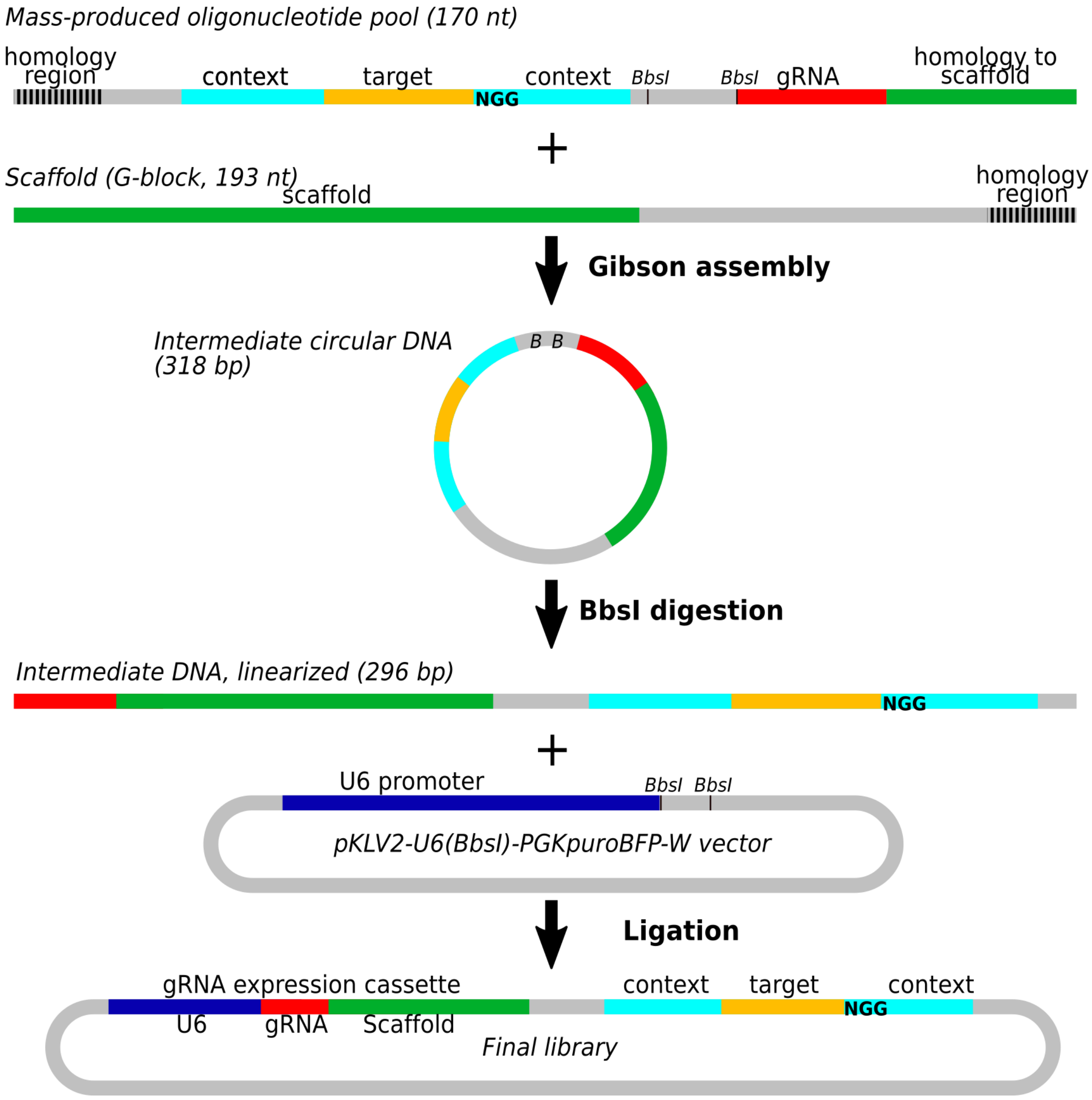
Supporting Information

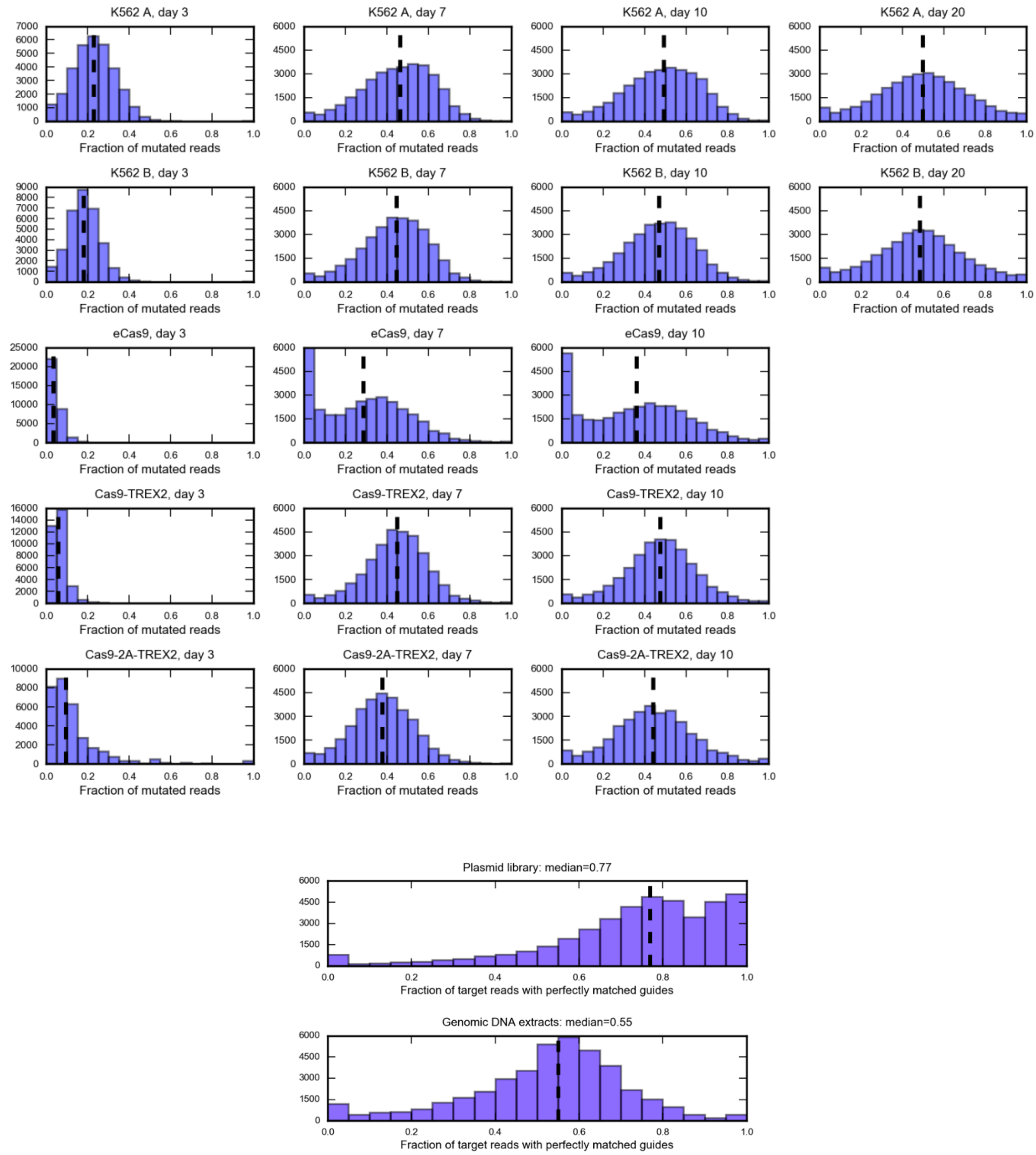
Supplementary Figure 1. Construct design and cloning approach



Library cloning started by PCR amplification of the 170 nt oligonucleotide pool of designed sequences encoding gRNA and target sequence, separated by a spacer harbouring two BbsI restriction sites. Gibson assembly was employed to fuse the amplified pool to a 193 nt G- block fragment encoding either a conventional or improved version of the gRNA scaffold and a spacer. The resulting 318 bp circular DNA was linearised with BbsI and the 296 bp linear product was ligated into scaffold-less pKLV2-U6(BbsI)-PKGpuro2ABFP-W, to produce the complete library constructs encoding a functional gRNA expression cassette and its target sequence.

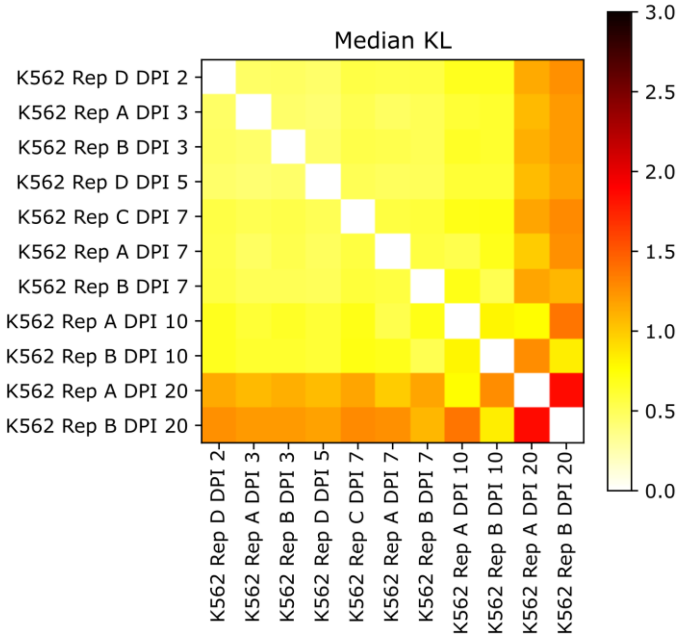
Nature Biotechnology: doi:10.1038/nbt.4317

Supplementary Figure 2. Fraction of edited targets during outgrowth



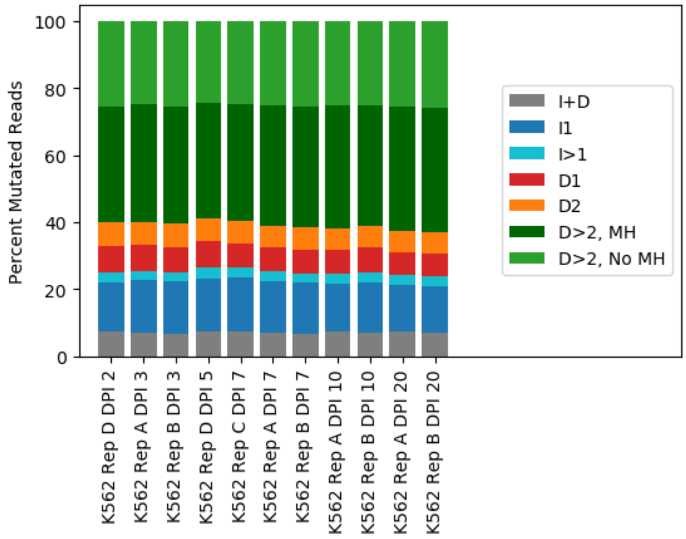
A. Editing rates are saturated after seven days of outgrowth. Number of gRNA-target pairs (y-axis) with a fraction of mutated reads (x-axis) across three, seven, 10, and 20 days of outgrowth (columns) for two replicates of K562 cell lines, and different effector proteins (rows). Dashed line: median fraction of mutated reads. Overall low rates of cutting (~50%) can be explained by synthesis errors and target switching due to viral recombination (B). B. Frequency (y-axis) of fraction of paired end reads that cover a guide RNA sequence on one end, and the correct target sequence on the other without any mismatches (x-axis). Top panel: sequencing results from the plasmid library. Bottom panel: sequencing results from genomic DNA extracts. Dashed line: median fraction of pristine reads. To obtain these data, we re-sequenced both the plasmid libary and a set of the other samples using a different pair of priming sites to capture both gRNA and its target barcode.

Supplementary Figure 3. Mutational profiles are consistent across early time points.



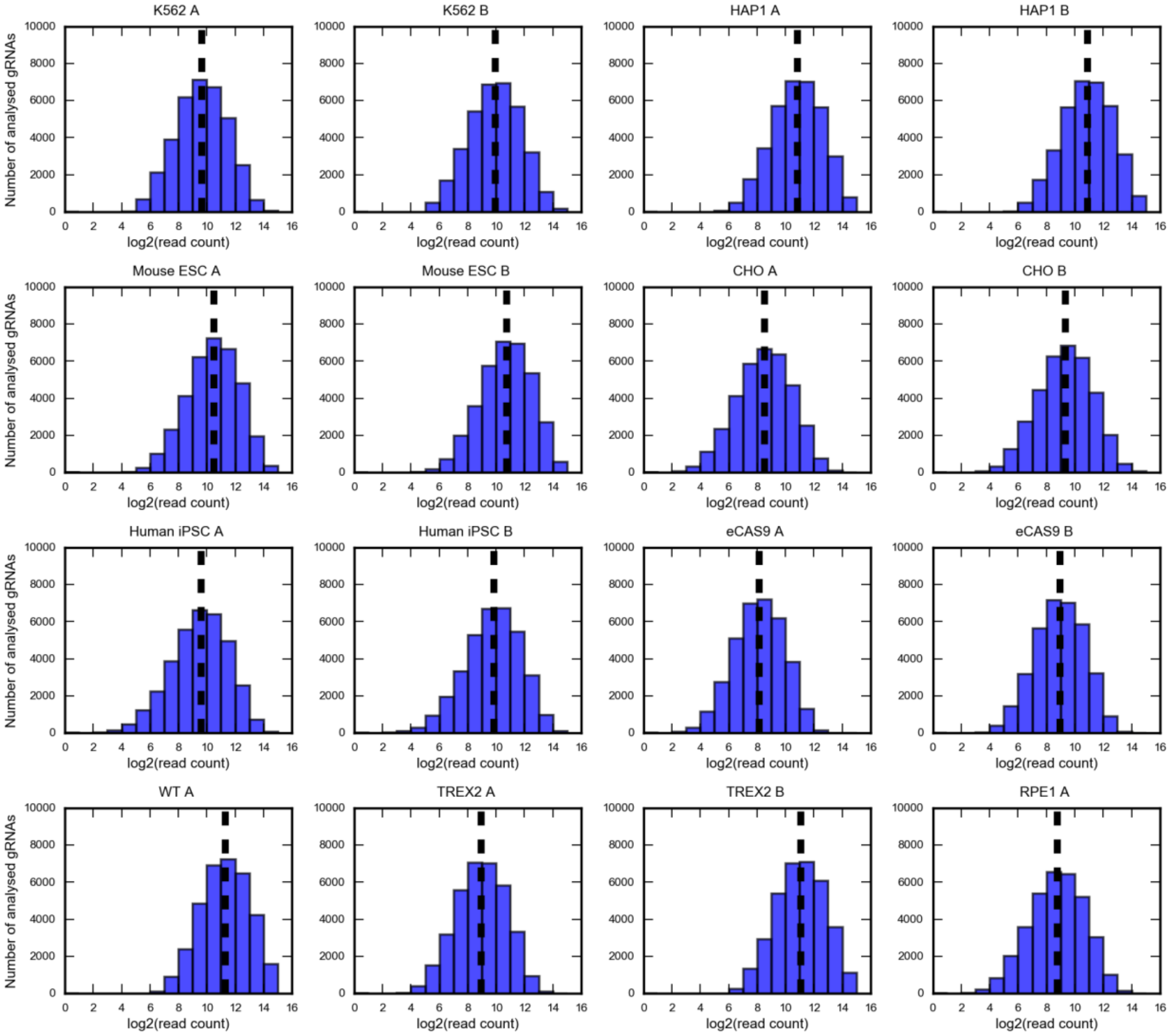
Median symmetric Kullback Leibler divergence between repair profiles (black to white color range, as in Figure 2B) at different time points (DPI=Days Post Infection) for K562 replicates (“A”, “B”, “C”, “D”) of 27,905 gRNAs from the “Explorative gRNA-Target” set (x and y axis). Mutational profiles are consistent both within and between replicates until DPI10, after which they diverge due to clonal evolution in culture.

Supplementary Figure 4. Proportion of mutated reads per indel type does not change substantially over time or across replicates



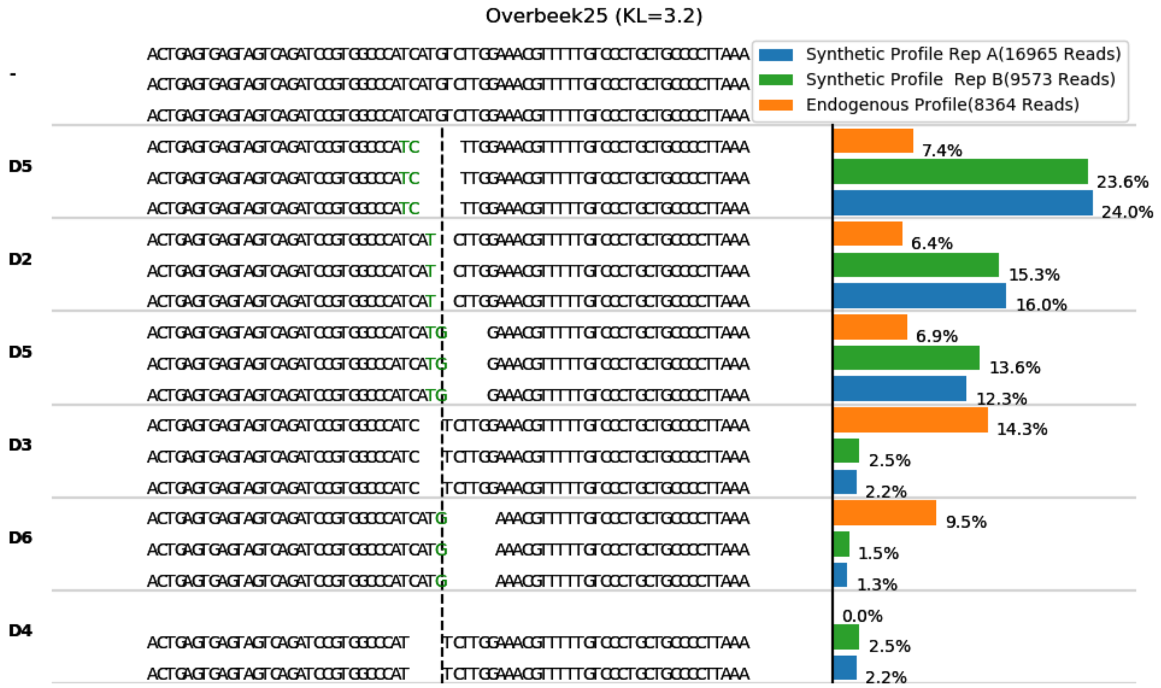
Frequency of different types of editing outcomes (y-axis; colors as 3B) for K562 replicates averaged across 27,905 gRNAs from the “Explorative gRNA-Target” set at various days post-infection (DPI). Samples as in Supplementary Figure 3; colors as in Figure 2B.

Supplementary Figure 5. Coverage across experiments



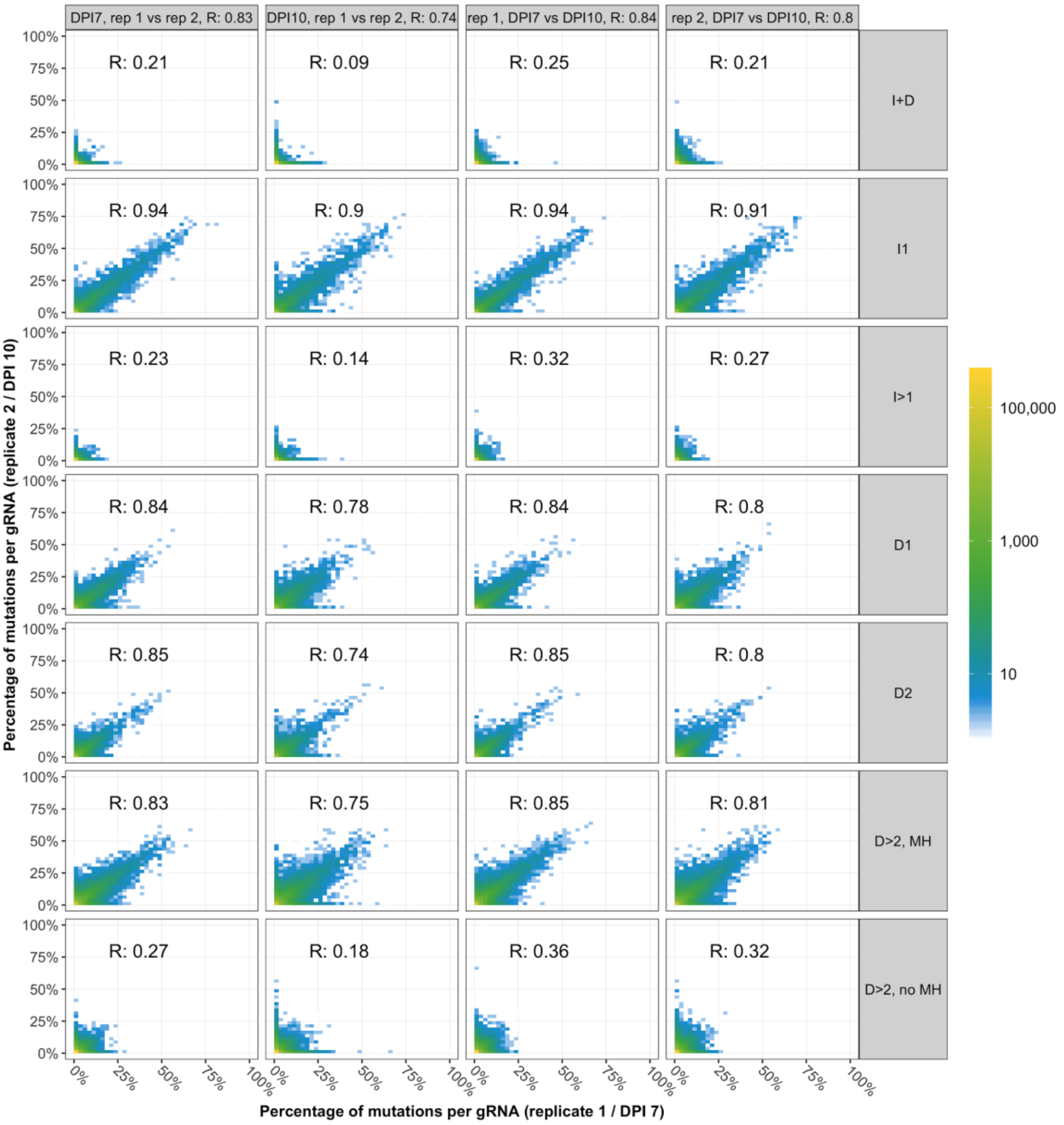
Number of gRNA-target pairs (y-axis) with a given sequencing coverage (x-axis; log2-scale) in the studied cell lines, effector proteins, and replicates. Dashed line: median coverage.

Supplementary Figure 6. Comparison of endogenous and synthetic repair profiles for van Overbeek 25 gRNA



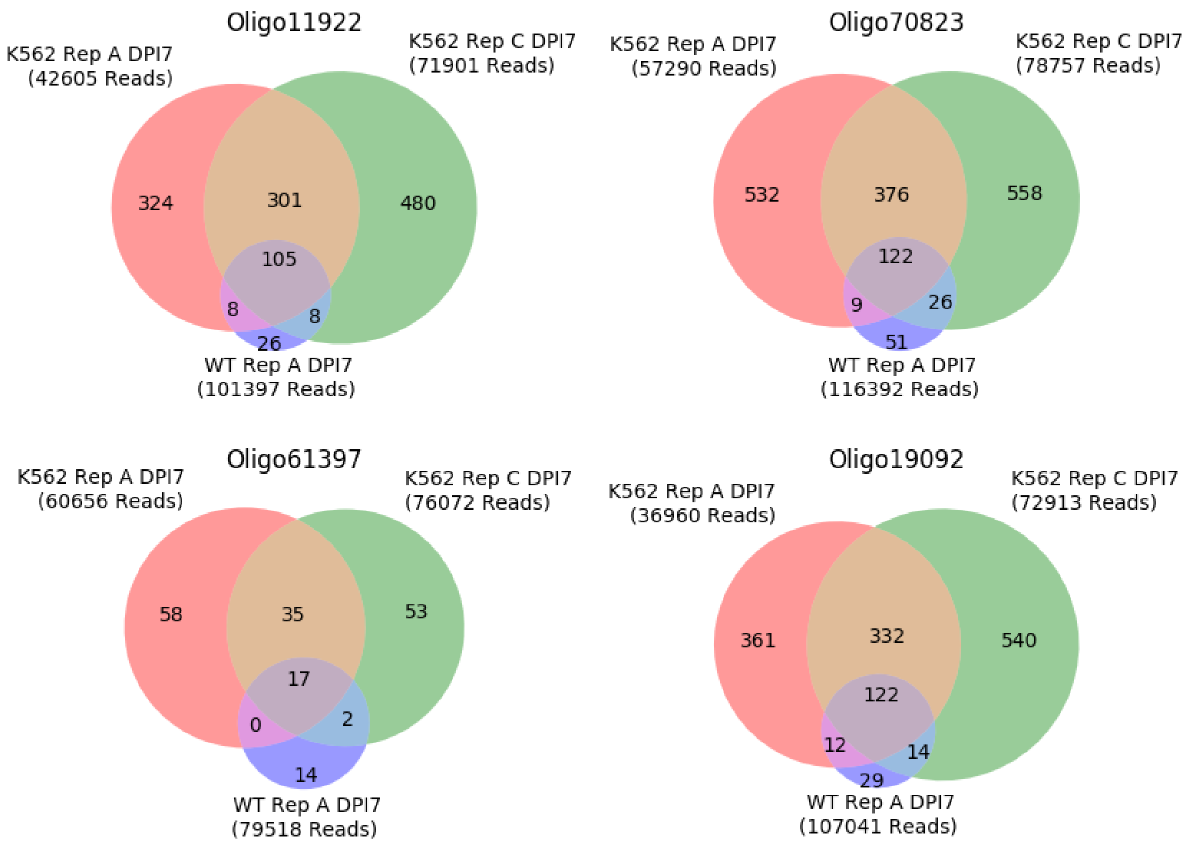
Measured repair profile reproducibility for the outlier, Overbeek 25, gRNA-target pair. DNA sequence of the target (top) is edited to produce a range of outcomes in two synthetic replicates (green, blue bars) and one endogenous measurement (orange bars). The proportions (x-axis) of the four largest mutational outcomes (e.g. “D3” - deletion of three base pairs, “I1” - insertion of one base pair, etc.; y-axis) is consistent between the experiments. Stretches of microhomology (green) and inserted sequences (red) are highlighted at the cut site (dashed vertical line).

Supplementary Figure 7. Reproducibility of indel frequencies per gRNA



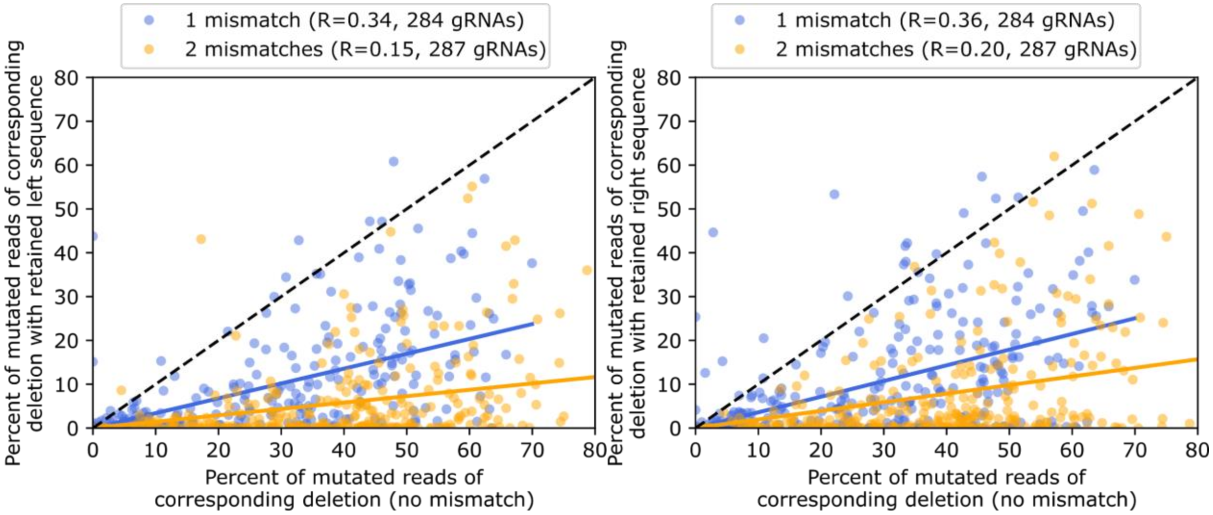
Number of indels (color) that compose increasing percentage of all mutations observed for their gRNA in sample 1 (x-axis) vs sample 2 (y-axis) depending on indel class (rows). Columns compare seven days post infection (DPI7) replicate 1 (R1) vs R2 (first column), DPI10 R1 vs R2 (second column), DPI7 R1 vs DPI10 R1 (third column), and DPI7 R2 vs DPI10 R2 (last column), with Pearson's R reported in the panel. Mutations from 6,568 gRNAs are represented in the figure.

Supplementary Figure 8. Low-frequency alleles are unlikely to be sequencing artifacts



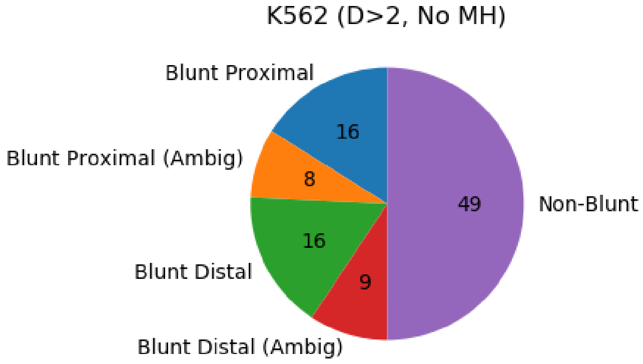
Number of distinct indels called for two different K562 replicates, Rep A (800x coverage) and Rep C (1600x coverage) and a control experiment carried out in wild type cells (WT, Cas9-, 800x coverage) for four oligonucleotides with very high sequencing coverage. The indels common to K562 and K562 WT (no Cas9) are mutations introduced during oligo synthesis or library construction and are removed from consideration in the indel calling code, unless they have significantly higher frequency than in the no Cas9 sample. The small number of mutations in the WT samples that are not observed in the K562 samples (blue) are indicative of the likely number of mutations called due to sequencing artefacts or low frequency oligo-synthesis mutations. This is a small fraction compared to the number of infrequent indels seen in K562 (red, green), which we can conclude are likely caused by Cas9.

Supplementary Figure 9. No bias in the side of sequence selected for microhomology-mediated end joining outcomes



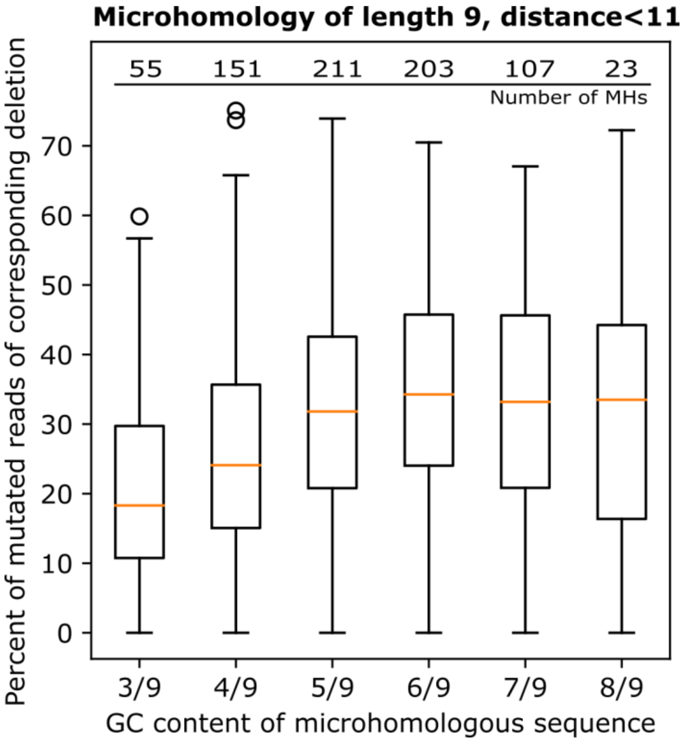
Mutations in microhomology sequence reduce repair outcome frequency, but corresponding deletions are still present. For matched pairs of guides, with and without mutations in the microhomologous sequence, the fraction of mutated reads associated with the particular microhomology (y-axis) is smaller than without mismatches (x-axis) for most gRNAs (markers; blue: one mismatch, yellow: two mismatches). The rates of repair are not different depending on whether sequence was retained from PAM-distal (left panel) or PAM-proximal (right panel) side of the cut. Pearson's R is reported in the legend.

Supplementary Figure 10. No bias in the side of deletions for nonhomologous end joining outcomes



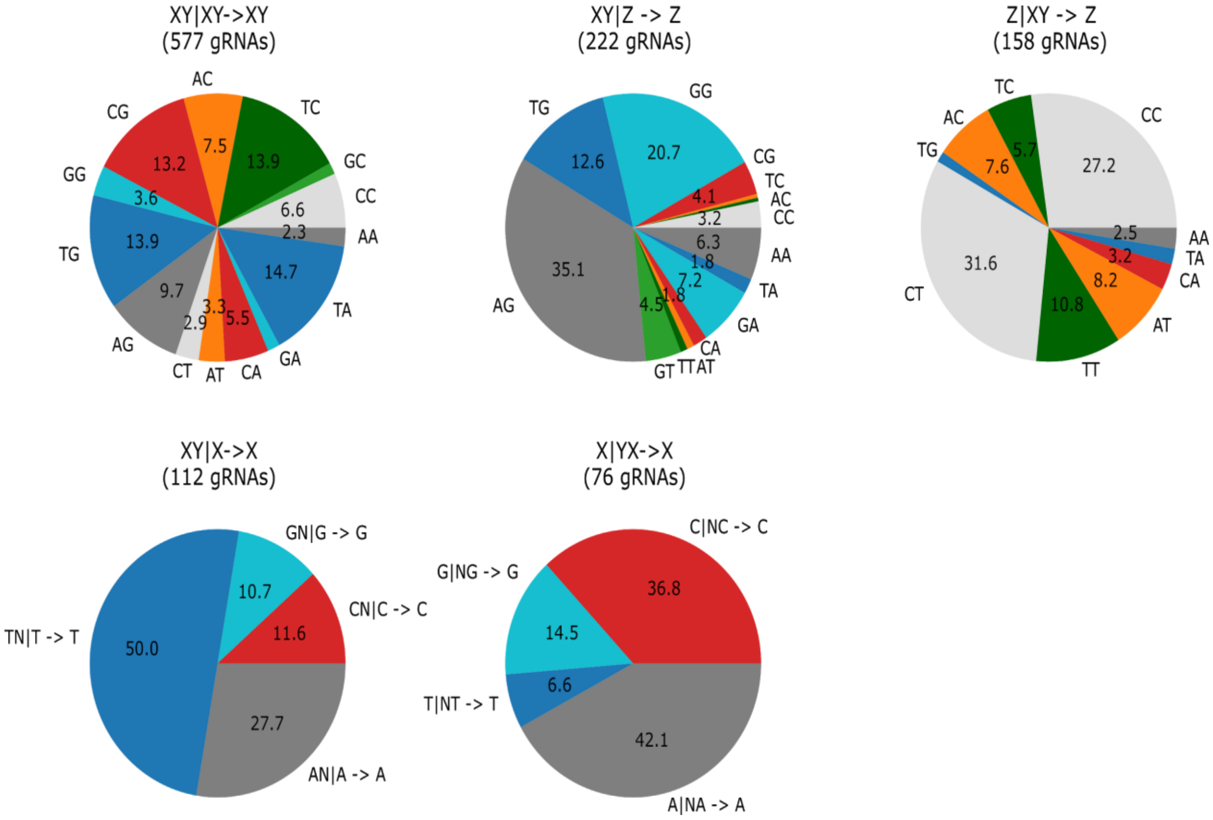
Average percent of alternative outcomes for large deletions without microhomology per-gRNA for 27,905 gRNAs from the “Explorative gRNA-Target” set (Methods). “Blunt” refers to deletions that occur exclusively on one side of the cut site and end precisely at the cut site. Distal and proximal refer to which side of the cut-site the deletion is on with respect to the PAM. “Ambig” refers to those deletions that could not be definitively assigned to an exact location (usually due to repeat nucleotides or microhomology of length 1) but could under at least one interpretation of their location be considered blunt.

Supplementary Figure 11. G+C content of the microhomologous sequence influences frequency of microhomology-mediated repair



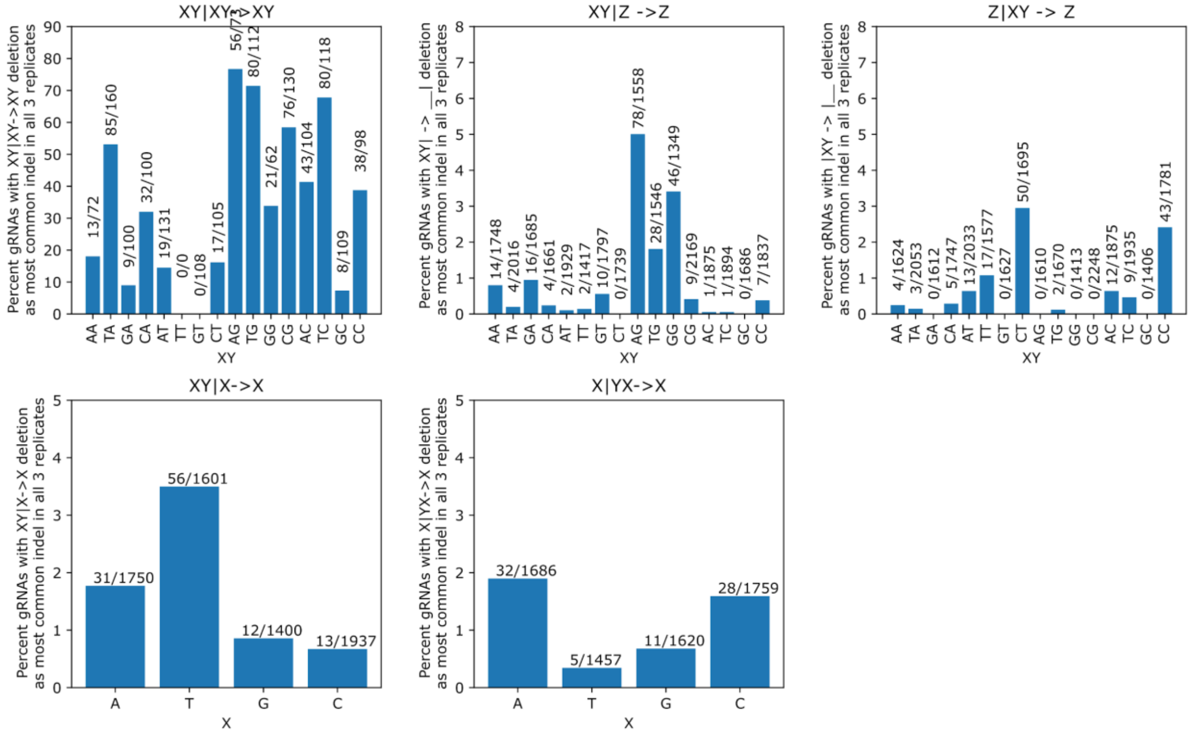
GC content influences microhomology-mediated repair fidelity. Percent gRNA reads with length 9 microhomology-mediated deletion (y-axis; boxes median and quartiles, whiskers 5% and 95%) across a range of GC contents (x-axis). 750 microhomology-pairs from 674 gRNAs total.

Supplementary Figure 12. Deletions of size 2 demonstrate strong biases in their preferred sequence characteristics, depending on presence of size 1 and 2 microhomology



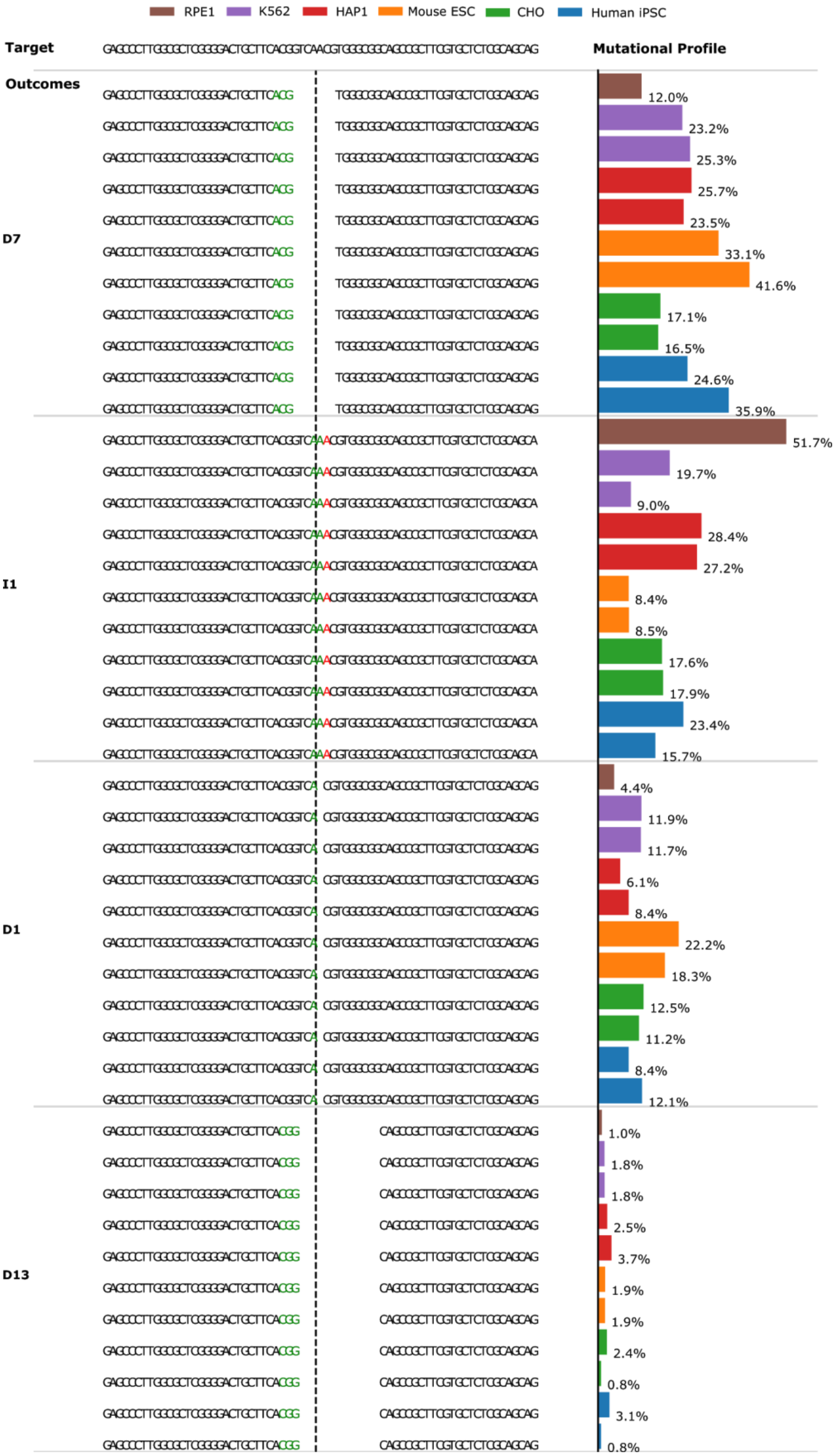
Nucleotides at the cut site bias the frequency of dominant dinucleotide deletion outcomes. Percent of dominant outcomes (area of wedge) with given sequence pattern surrounding the cut site (colors) for different types of deletions (panels; types as in Figure 4H). X, Y, Z, W, N - any nucleotide.

Supplementary Figure 13. Sequence at the cut site biases the rate of deletions of size 2, depending on repeating single and double nucleotides

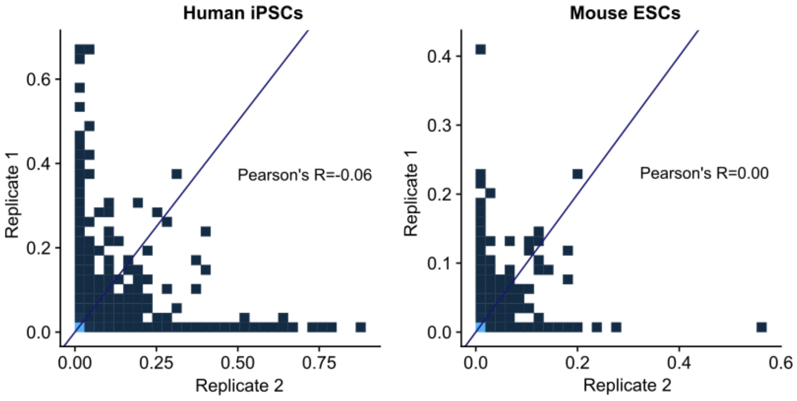


Nucleotides at the cut site bias the rate of dominance of dinucleotide deletion. Percent of gRNAs for which a dinucleotide deletion is dominant (y axis) with given sequence pattern surrounding the cut site (x axis) for different types of deletions (panels; types as in Figure 4H and S12). X, Y, Z, W, N - any nucleotide.

Supplementary Figure 14. Example of profile measured across different cell lines

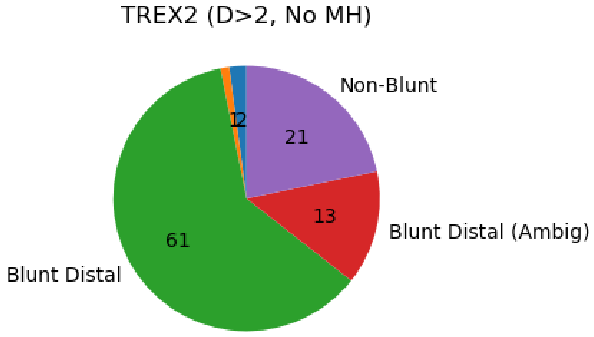


Supplementary Figure 15. Lack of reproducibility in large insertions in human and mouse stem cells



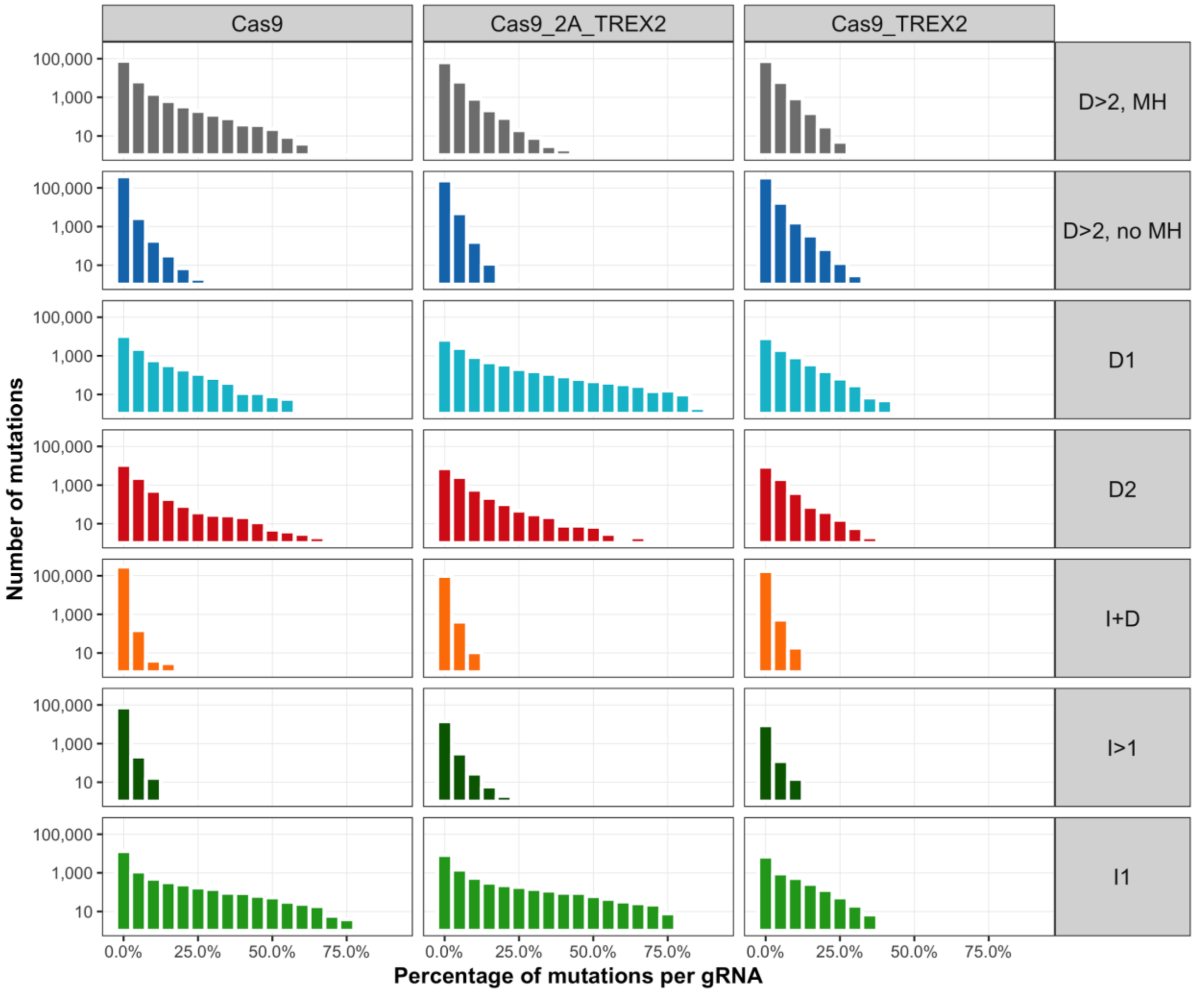
Frequency of individual insertions of size at least 2 (x and y axis) in two replicates of human iPSCs (left) and mouse ESCs (right) in bins of 2%, colored by number of mutations with the corresponding frequency. A very small number of insertions are observed in both replicates. Mutations are collected from 6,568 gRNAs.

Supplementary Figure 16. Repair outcomes from Cas9-TREX2 fusion favor blunt end joins with the deletion on the PAM-distal side



Average per-gRNA percent of outcomes for deletions of at least 3nt without microhomology for 27,905 gRNAs from the “Explorative gRNA-Target” set (Methods). “Blunt” refers to deletions that occur exclusively on one side of the cut site and end precisely at the cut site. Distal and proximal refer to which side of the cut-site the deletion is on with respect to the PAM. “Ambig” refers to those deletions that could not be definitively assigned to an exact location (usually due to repeat nucleotides or microhomology of length 1) but could under at least one interpretation of their location be considered blunt.

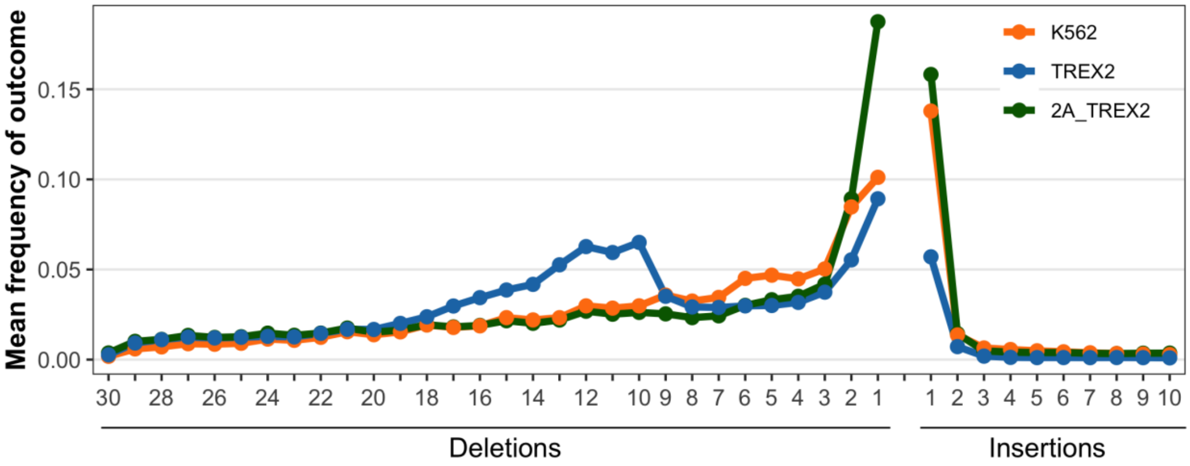
Supplementary Figure 17. TREX2 overexpression effects on indel class frequencies



Per-gRNA event frequencies change upon TREX2 expression. Number of individual indels (y-axis) as a percentage of all mutations observed for their gRNA (x-axis) separated by mutation class (rows), and Cas9/TREX2 construct (columns).

Nature Biotechnology: doi:10.1038/nbt.4317

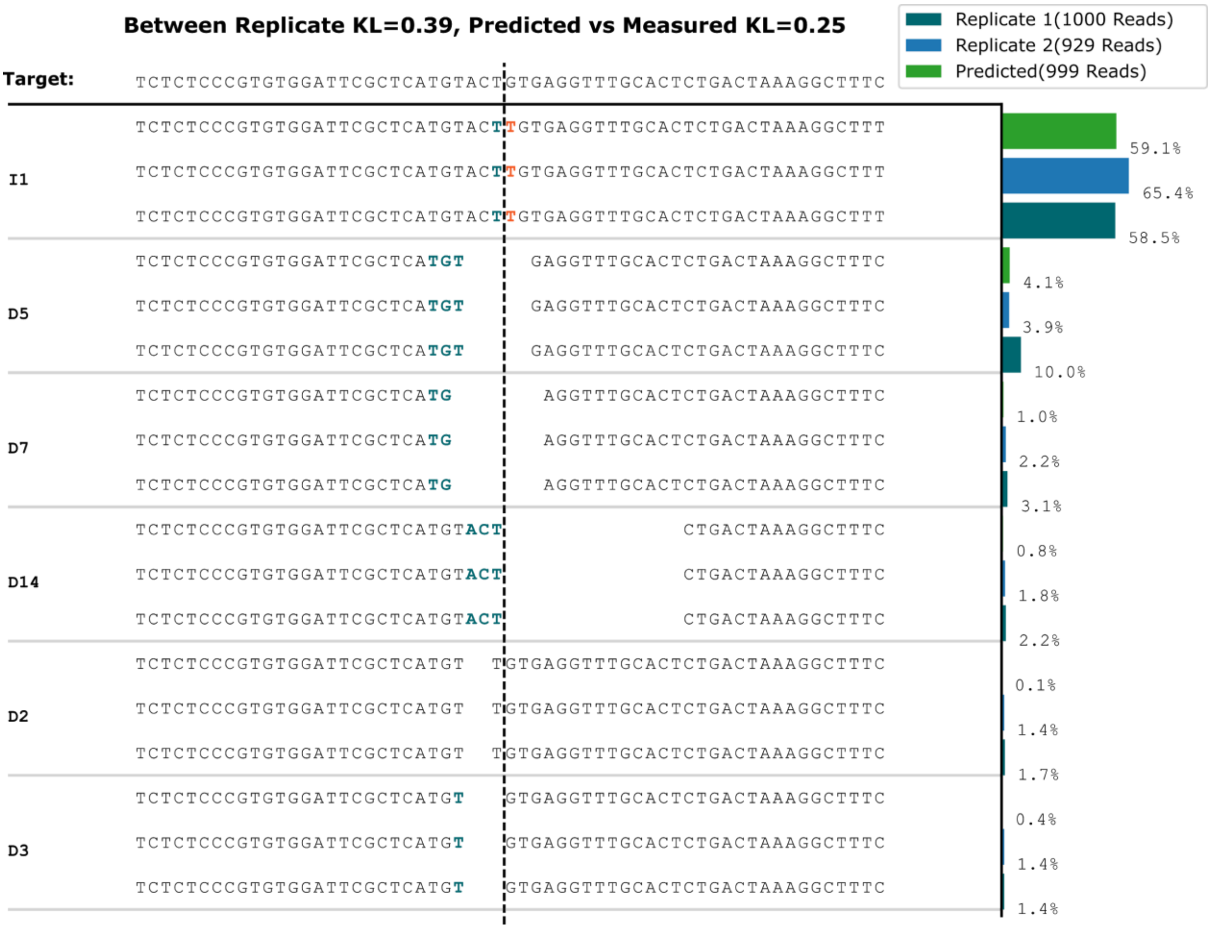
Supplementary Figure 18. Cas9-2A-TREX2 has a different influence on repair outcomes compared to Cas9-TREX2



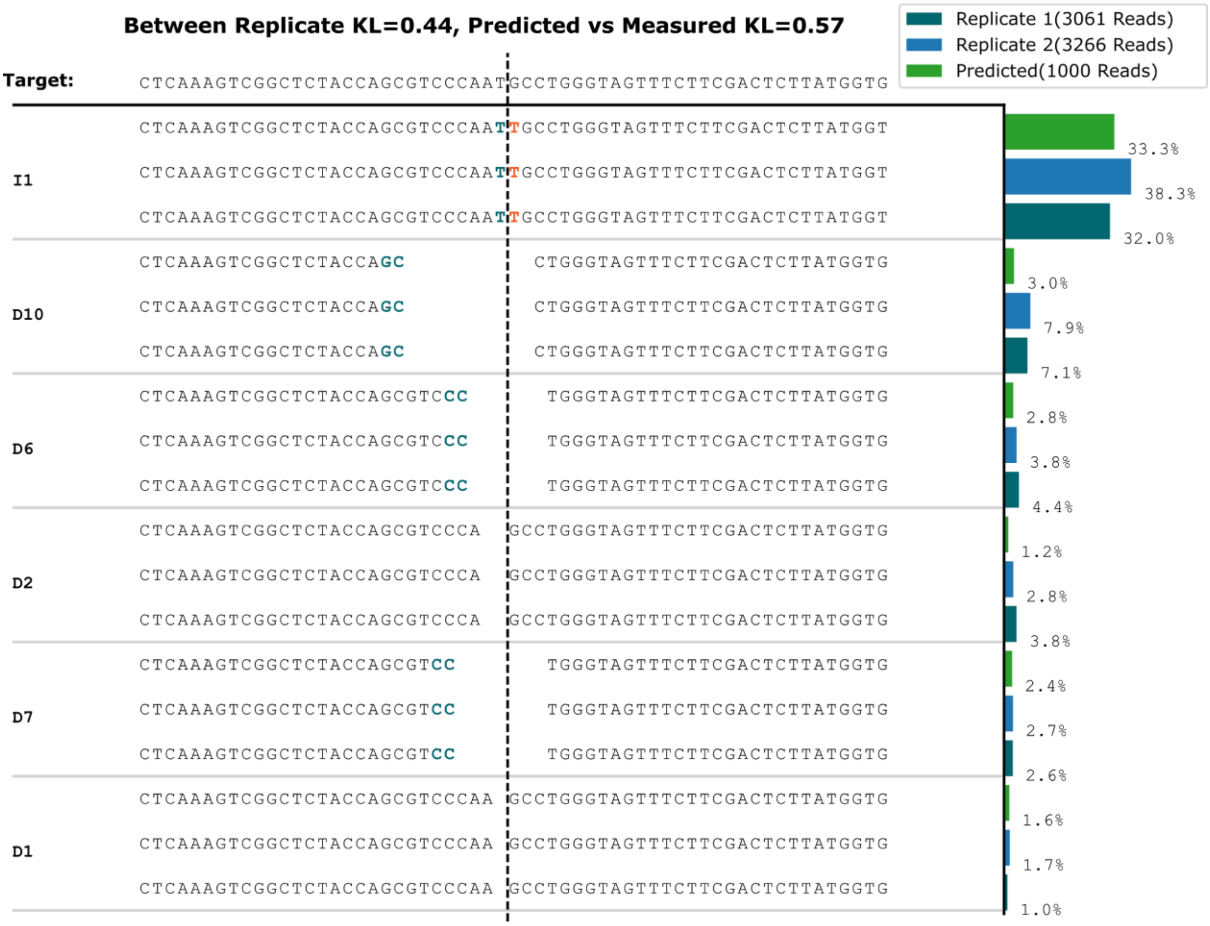
The mean frequency (y-axis) of deletion or insertion size (x-axis) across genomic sequence targets for three alternative Cas9 effector constructs (colors).

Nature Biotechnology: doi:10.1038/nbt.4317

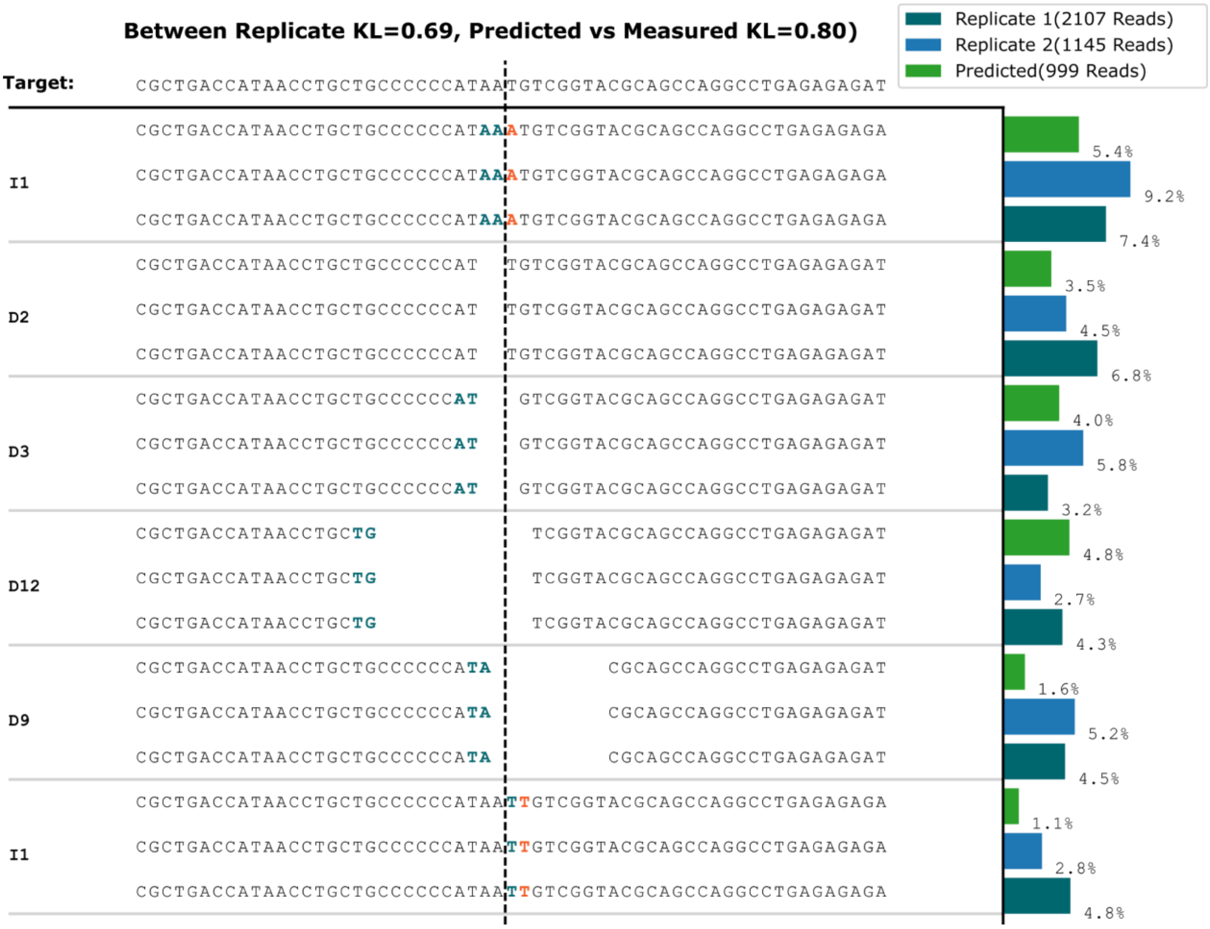
Supplementary Figure 19a. Example predicted mutational profile with KL = 0.25 (lower whisker in Figure 6b) for predicted vs. measured; two replicate measurements in K562 also shown, for which the KL is 0.39.



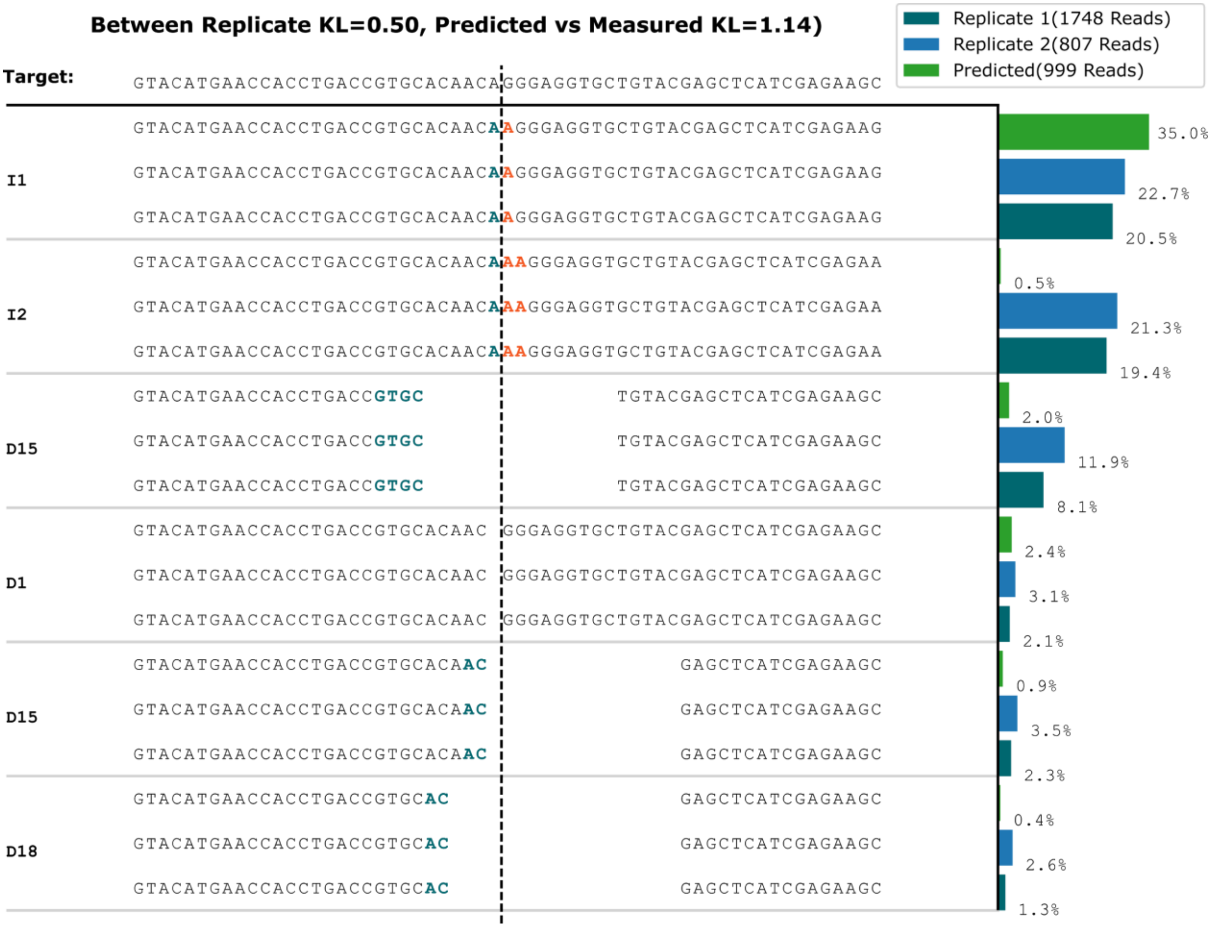
Supplementary Figure 19b. Example predicted mutational profile with KL = 0.57 (lower quartile line in Figure 6b) for predicted vs. measured; two replicate measurements in K562 also shown, for which the KL is 0.44.



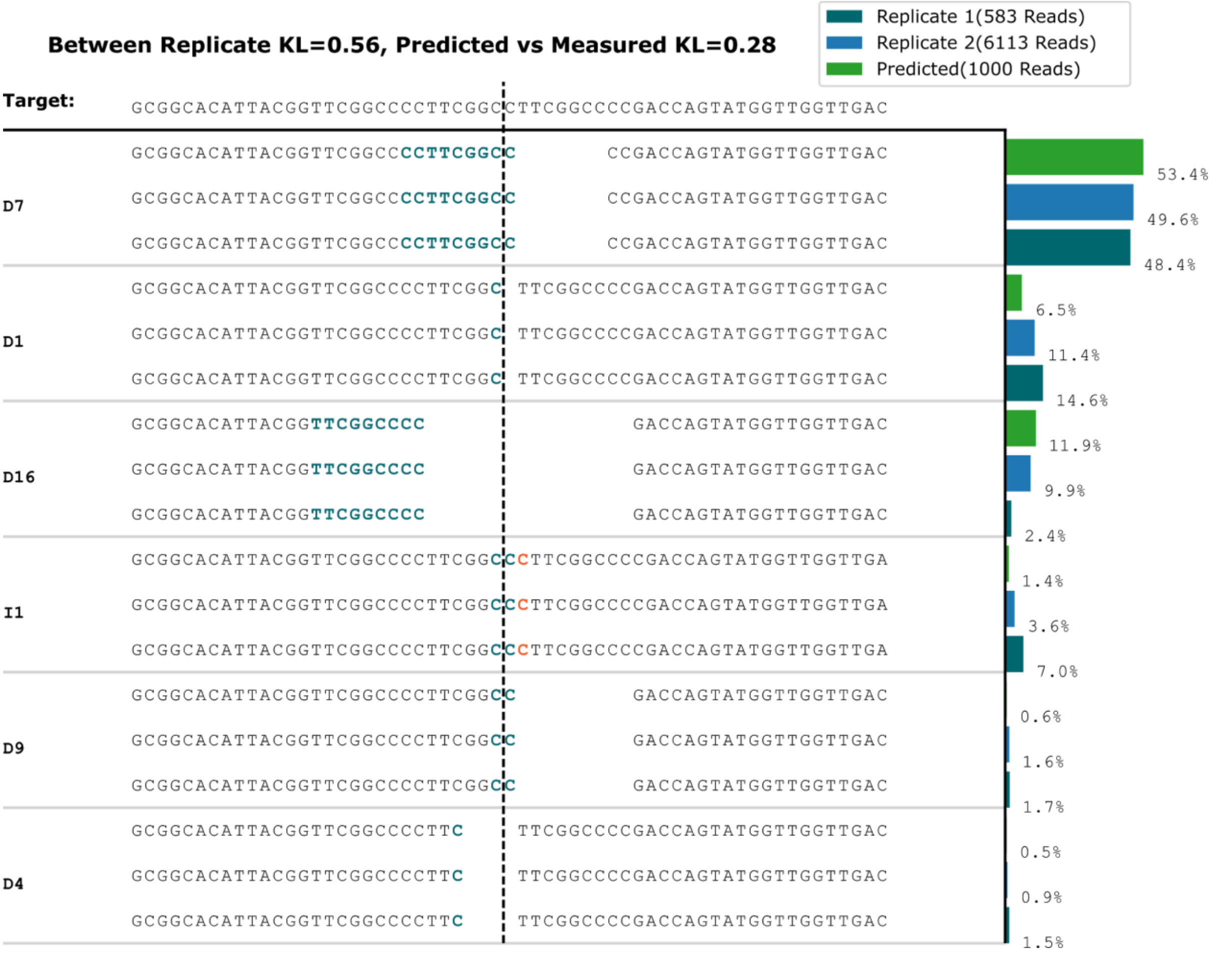
Supplementary Figure 19c. Example predicted mutational profile with KL = 0.80 (upper quartile line in Figure 6b) for predicted vs. measured; two replicate measurements in K562 also shown, for which the KL is 0.69.



Supplementary Figure 19d. Example predicted mutational profile with KL = 1.14 (upper whisker in Figure 6b) for predicted vs. measured; two replicate measurements in K562 also shown, for which the KL is 0.50

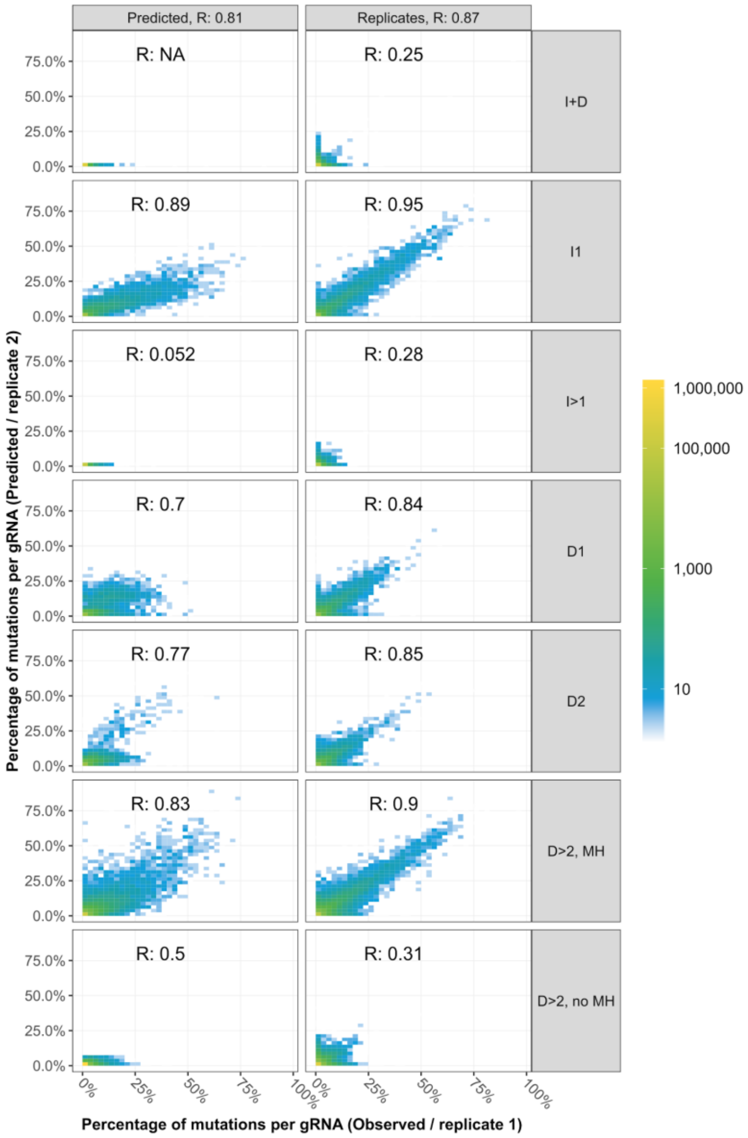


Supplementary Figure 19e. Example predicted mutational profile with competing two microhomologies of length 9 (D7 and D16) and KL = 0.28 for predicted vs. measured; two replicate measurements in K562 also shown, for which the KL is 0.56



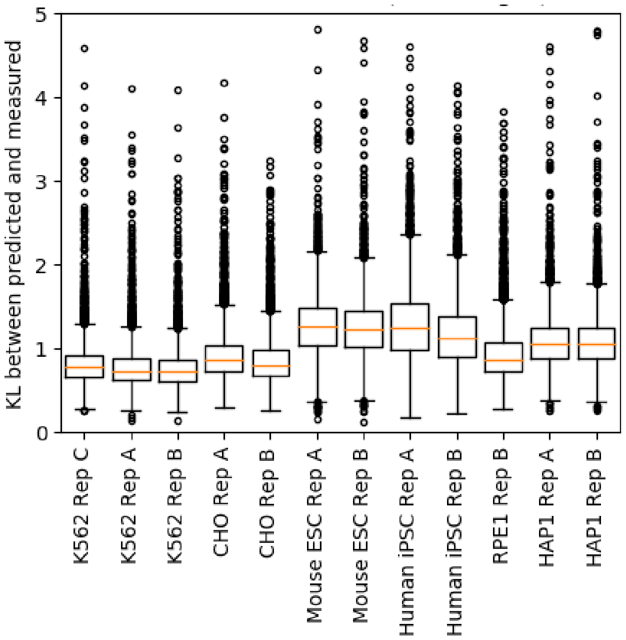
Nature Biotechnology: doi:10.1038/nbt.4317

Supplementary Figure 20. Comparison of allele frequencies from predictions and replicates



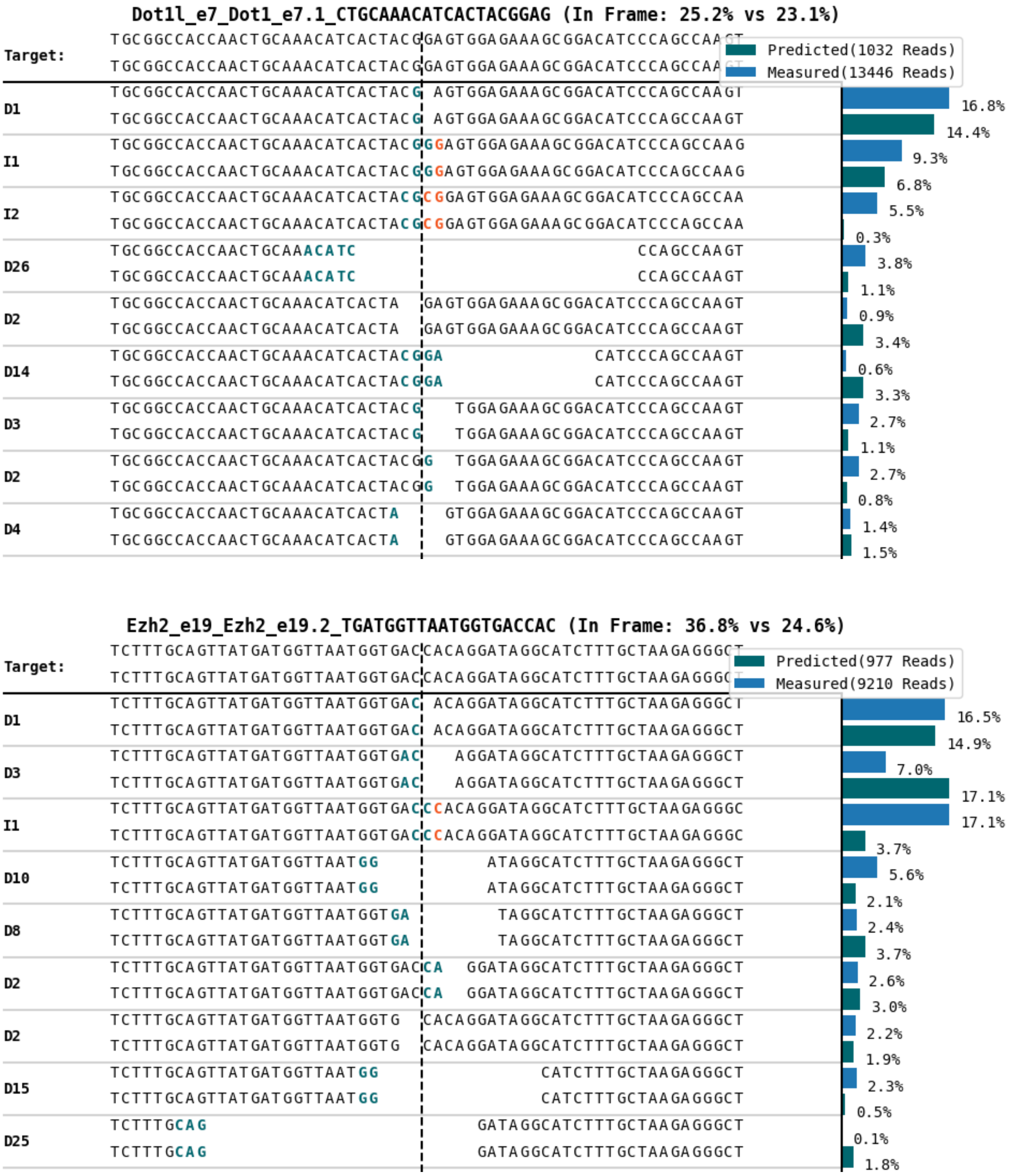
Density of frequency of mutations estimated by two approaches. Left panel: predicted (y-axis) vs combined measurement (x-axis); right panel: measurement replicate 1 (x-axis) vs measurement replicate 2 (y-axis), both categorised by mutation type (rows), with Pearson’s correlation given. I+D mutations and I>2 mutations are not predicted by the model. Mutations are collected from 6,568 gRNAs.

Supplementary Figure 21. The performance of the predictor trained on K562 in other cell lines is related to their similarity to K562

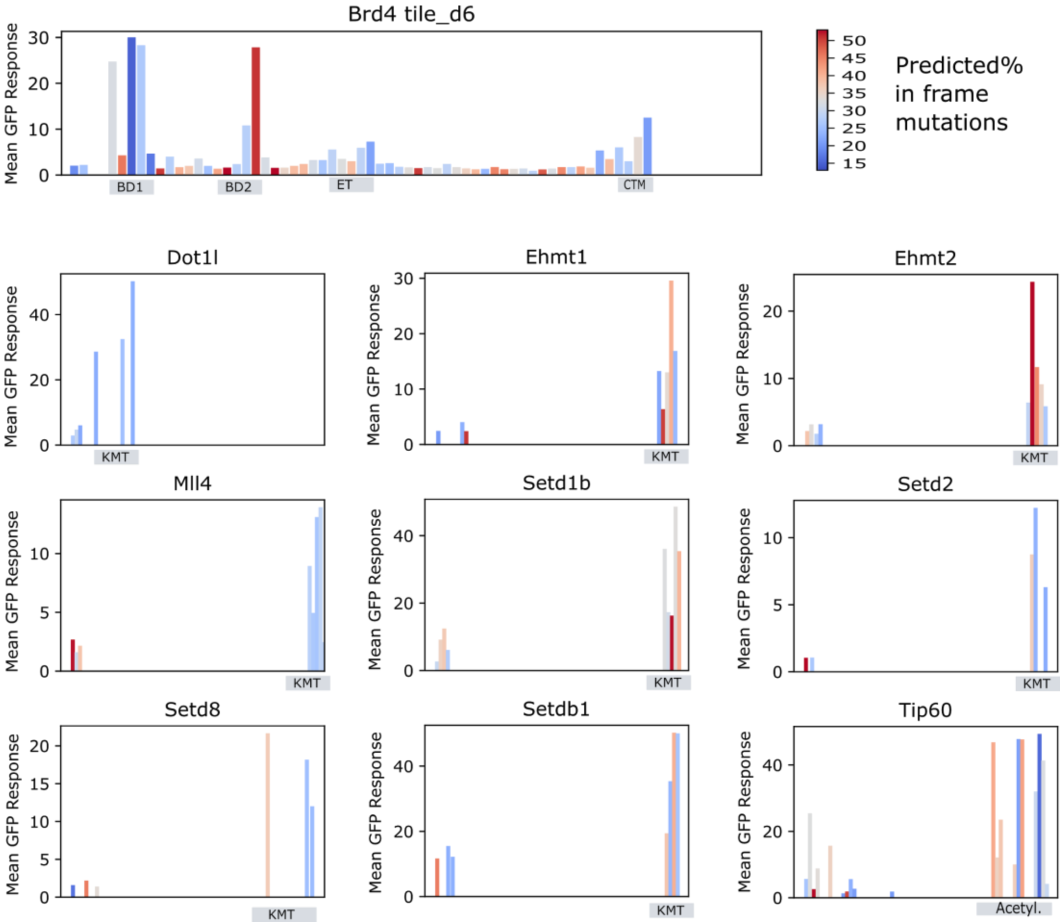


Box plot of symmetric KL divergence (y-axis; orange line: median; box edges: quartiles; whiskers: 5% and 95%) between our predictions, and the measured profiles in different cell lines (x-axis) on a subset of 4,722 validation gRNA-target pairs, filtered to contain at least 20 reads for all gRNAs tested across all samples. Performance is best for K562, and unsurprisingly declines more in cell lines for which we observed more dissimilar profiles in Figure 5B.

Supplementary Figure 22. Example comparisons between gRNAs measured by Shi *et al.* 2015 in the murine MLL-AF9/NrasG12D acute myeloid leukemia cell line (RN2) and our predictions (trained on data from K562 cells).



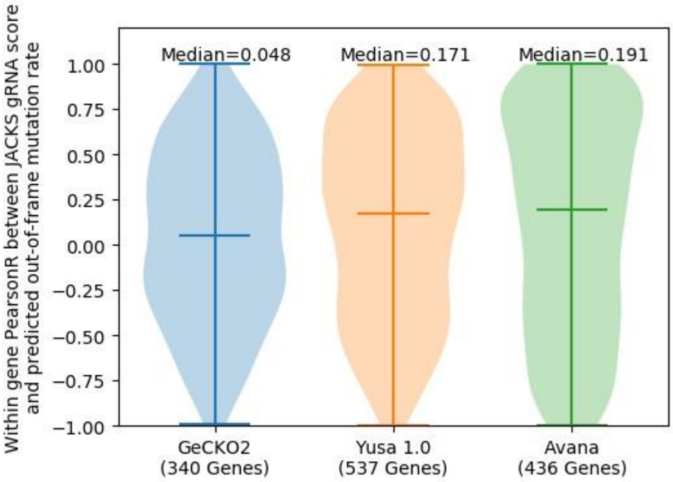
Supplementary Figure 23. Predicted rates of in-frame mutation do not explain variability in knockout effect in data from Shi *et al.*



FORECasT predicted in-frame mutation percentage (colours from blue (low) to red (high)) applied to data from Shi et al (Figure 2a,4b-j), which quantifies the phenotypic effect (y-axis) for gRNAs targeting different regions of the gene (x-axis; 5’ to 3’) for 10 genes (panels), with protein domains highlighted in grey rectangles. There is no obvious relationship between the heights of the bars and their colour, either within domain or outside domain, and in the 5’ or 3’ end of the gene.

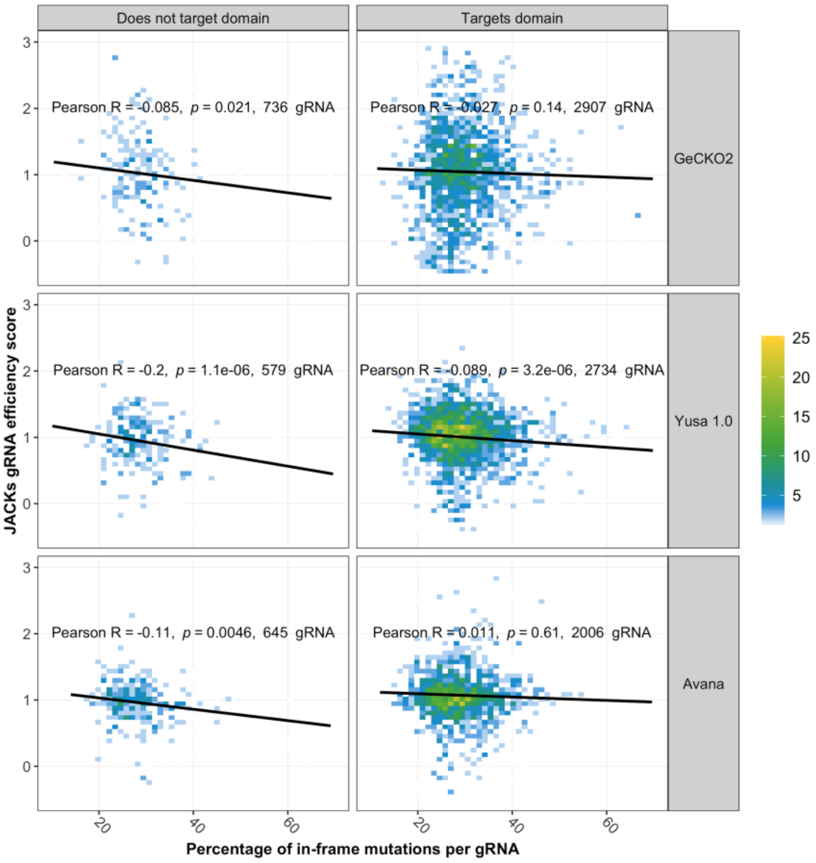
Nature Biotechnology: doi:10.1038/nbt.4317

Supplementary Figure 24. Predicted out-of-frame fraction is positively correlated with gRNA efficacy



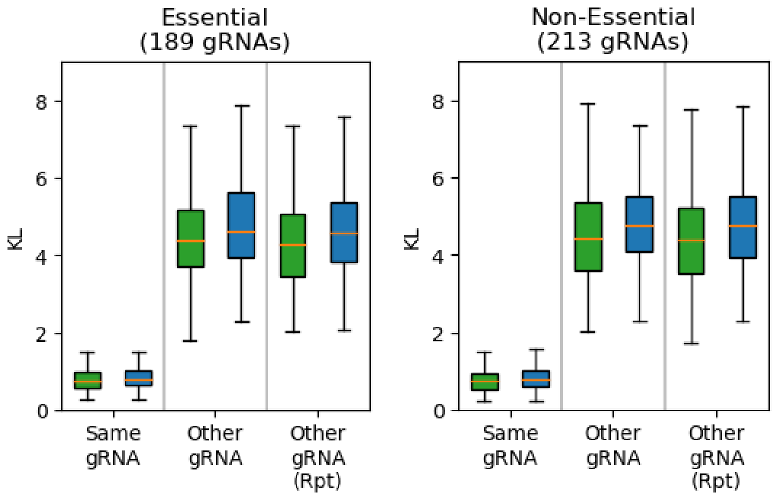
Density (x-axis) of Pearson’s correlation coefficients (y-axis) calculated between JACKS-inferred gRNA efficacy (Allen et al. 2018) and predicted out-of-frame mutation rate for essential genes from three genome-wide libraries (colors; x-axis panels): (Meyers et al. 2017) (Avana library), (Tzelepis et al. 2016) (Yusa v1.0 library), and Aguirre et al 2016 (Aguirre et al. 2016) (GeCKO v2 library). Median values marked on the plot and denoted above the violins (central marker); range given as external line markers. Number of genes making up the density is provided below the library labels.

Supplementary Figure 25. Fraction of out-of-frame mutations is more important for gRNA efficacy when gRNAs target outside protein domains.



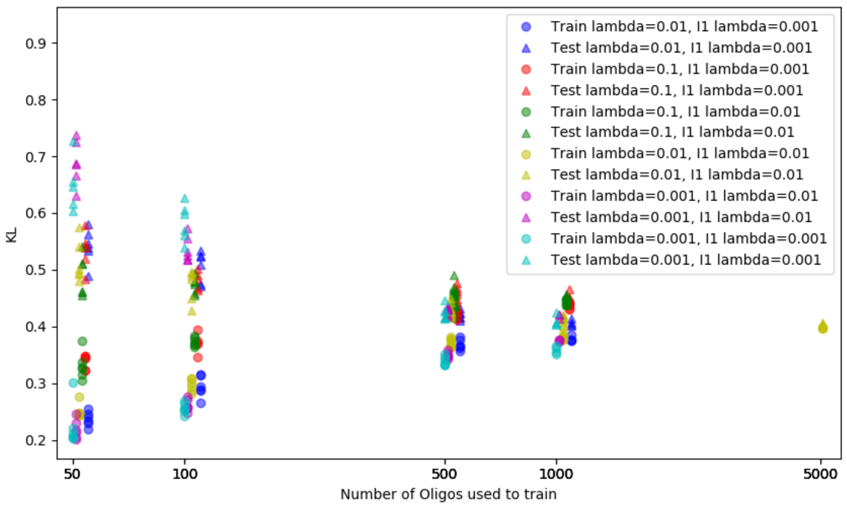
JACKS-inferred gRNA efficacy (y-axis) and predicted out-of-frame mutation rate (x-axis) for gRNAs targeting essential genes from three genome-wide libraries (rows) stratified by gRNAs targeting outside (left column) and inside (right column) protein domains. Pearson's correlation with corresponding p-value, linear model fit, and dataset size is given on the plot.

Supplementary Figure 26. Gene essentiality has no impact on mutational profile reproducibility



As Figure 2C, but restricted to gRNAs targeting targeting essential genes (left panel; between replicate median KL=0.73) and non-essential genes (right panel; between replicate median KL=0.74) as defined by (Hart et al. 2017). Box plots: orange median line, quartiles for box edges, 95% whiskers.

Supplementary Figure 27. Exploration of model hyperparameters and training set size



Optimizing FORECasT model hyperparameters and training set size in using random samples from the Explorative gRNA-target set (disjoint set from hold-out guides used in validation). The selected model used all regularization parameters set to 0.01 and trained on 5000 gRNAs (yellow, right-most markers). At this point the average KL values of the training and test set have converged, indicating an absence of over-fitting. The KL values used here are asymmetric and so give smaller values than those reported in all other measurements.

Supplementary Table 1. Screen conditions

Cell line	Cells ($x10^6$)	Multiplicity of infection	Coverage (cells per construct)	Replicates	Time points sequenced (days post infection)
K562-Cas9	70	0.6	800	2	3-7-10-20
K562-Cas9	140	0.6	1600	2	3-7, 2-5
K562-Cas9 with conventional scaffold	16, 16, 32	0.5	800,800,1600	3	7
K562-eCas9	70	0.6	800	2	3-7-10
K562-Cas9-TREX2	70	0.6	800	2	3-7-10
K562-Cas9-2A-TREX2	70	0.6	800	2	3-7-10
RPE-1-Cas9	52	0.5	500	1	7
HAP1-Cas9	83	0.5	800	2	7
CHO-Cas9	83	0.5	800	2	7
iPSC-Cas9	83	0.5	800	2	7
E14TG2a-Cas9	83	0.5	800	2	7
K562-No Cas9	70	0.6	800	1	7
K562-No Cas9 with conventional scaffold	16	0.5	800	1	7

Supplementary Table 2. Average KL divergence between synthetic and endogenous measurements stratified by ChromHMM classification

ChromHMM	Median KL	Number of endogenous target regions
CtcfO	0.79	3
DnaseD	0.95	3
ElonW	1.26	1
Enh	2.73	1
EnhF	1.00	2
EnhW	0.74	1
EnhWF	0.75	2
Gen3’	1.23	5
Gen5’	1.12	7
Low	0.99	9
PromP	1.79	2
Quies	0.99	9
Repr	1.11	5
ReprW	1.25	7
Tss, or “Active Promoter”	1.16	20
“Enhancer” (EnhWF, EnhF, Enh, EnhW)	0.82	6
“Transcription” (Elon, ElonW, Gen3, Gen5, Pol2)	1.14	13
“Repressed” (Quies, Repr, ReprW)	1.04	21

Supplementary Table 3. Strongest feature values for prediction; note that insertion features tend to feature more strongly because there are fewer of them

Feature Symbol	θ value/s	Description
I1Rpt	0.744	Single nucleotide insertion repeating the PAM distal nucleotide adjacent to the cut site
IL-1--1, IL-2--2	0.996, 0.576	Insertion at the cut site
PW_CS0_NT=G_vs_I1Rpt	0.473	I1Rpt and G at the PAM-proximal nucleotide adjacent to the cut site
PW_CS-1_NT=T_vs_I1Rpt	0.413	I1Rpt and T at the PAM-distal nucleotide adjacent to the cut site
PW_I1_T_vs_I1Rpt	0.408	Single nucleotide insertion repeating the PAM distal nucleotide adjacent to the cut site in which a T is inserted
PW_CS-2_NT=C_vs_I1Rpt	0.404	Single nucleotide insertion repeating the PAM distal nucleotide when there is a C in the next most distal nucleotide.
PW_CS2_NT=A_vs_I1Rpt	0.375	Single nucleotide insertion repeating the PAM distal nucleotide when there is an A next to the start of the PAM.
PW_No MH_vs_DL-1--1	0.341	No microhomology, blunt-end deletion on PAM-proximal side
PW_CS-1_NT=A_vs_I1Rpt	0.300	I1Rpt and A at the PAM-distal nucleotide adjacent to the cut site
PW_D>12_vs_DL>=0	0.294	Deletions greater than 12 and ambiguity of location in the left side of the deletion (indicates microhomology).
PW_I1_A_vs_I1Rpt	0.290	Single nucleotide insertion repeating the PAM distal nucleotide adjacent to the cut site in which an A is inserted
PW_R-1_NT=G_vs_DR0-0	0.244	Blunt deletion on the PAM-distal side of the cut site if there is a G as the first removed nucleotide closest to the cut site.
...		
>3000 other features		
...		
I1_G	-0.327	Single nucleotide insertion of a G
PW_D>12_vs_DR0-0	-0.388	Blunt-end deletion on PAM-distal side of cut of size greater than 12

Supplementary Table 4. Evaluation of predicted rates of in-frame mutation in 12 deep-sequenced gRNAs from Shi *et al.*

gRNA	Percent predicted in-frame mutations	Percent measured in-frame mutations	Predicted vs Measured KL	Percent deletions with size > 30 in measured profile
Brd4_e3.1	32.3	34.2	1.0	12.6
Brd4_e3.3	30.6	18.6	1.3	8.5
Brd4_e4.1	17.3	13.8	1.5	4.9
Dot1l_e1.1	31.5	29.4	0.78	10.0
Dot1l_e3.1	28.3	26.7	0.72	8.9
Dot1l_e7.1	25.2	23.1	1.2	13.9
Dot1l_e11.1	26.5	42.8	6.0	91.2
Ezh2_e2.2	22.7	18.1	1.1	13.9
Ezh2_e19.2	36.8	24.6	0.83	7.0
Smarca4_e2.1	34.7	22.4	1.8	3.67
Smarca4_e3.1	26.1	27.5	1.1	11.6
Smarca4_e16.1	30.6	37.5	1.0	15.8

Supplementary Table 5. Characteristics of explorative guide set

Length of Microhomology	Number of gRNA-Targets with microhomology at distance.		
	< 5	< 10	< 20
3	7923	15236	26141
4	2956	6459	15627
5	1143	2517	5509
6	591	1355	2732
7	394	935	1748
8	321	775	1372
9	280	674	1212
10	279	678	1144
11	313	712	1124
12	284	650	1026
13	267	648	1040
14	285	698	1106
15	259	649	987

Supplementary Table 6. Primer sequences (5′ > 3′)

Cloning of pKLV2-U6(BbsI)-PKGpuro2ABFP-W

P1	GGCAGCACTGCATAATTCTCTTAC
P2	CCTACCCGGTAGAATTGGATCCAAACGTGTCTTCTCGAAGACCC
P3	GTAAGAGAATTATGCAGTGCTGCC
P4	GGGTCTTCGAGAAGACACGTTTGGATCCAATTCTACCGGGTAGG

Amplification of oligo pools for library cloning

P5	GGAAACTACACTTGCCTGGC
P6	AACTTGCTATTTCTAGCTCTAAAAC
P7	GACGTCCAGAGCACAGATGG
P8	GCTGTTTCCAGCATAGCTCTTAAAC

Preparation of sequencing libraries

P9	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCTTGTGGAAAGGACGAAACA
P10	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGGAAACTACACTTGCCTGGC
P11	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGACGTCCAGAGCACAGATGG
P12	TCGGCATTCCTGCTGAACCGCTCTTCCGATCTACCCGGTAGAATTGGATCCAAAC
P13	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
P14*	CAAGCAGAAGACGGCATACGAGATN_{10}GAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Sequencing primers

P15	TCTTCCGATCTCTTGTGGAAAGGACGAAACACCG
P16	CTCTTCCGATCTGACGTCCAGAGCACAGATGG
P17	GCTCTTCCGATCTGGAAACTACACTTGCCTGGC
P18	CGCTCTTCCGATCTACCCGGTAGAATTGGATCCAAAC

*: N_{10}, index for multiplexed sequencing.