Predictable and precise template–free CRISPR editing of pathogenic variants

Max W. Shen1,2,12, Mandana Arbab3,4,5,12, Jonathan Y. Hsu6,7, Daniel Worstell8, Sannie J. Culbertson8, Olga Krabbe8,9, Christopher A. Cassa8,10, David R. Liu3,4,5,*, David K. Gifford2,6,10,11 & Richard I. Sherwood8,9,12

Following Cas9 cleavage, DNA repair without a donor template is generally considered stochastic, heterogeneous and impractical beyond gene disruption. Here, we show that template–free Cas9 editing is predictable and capable of precise repair to a predicted genotype, enabling correction of disease–associated mutations in humans. We constructed a library of 2,000 Cas9 guide RNAs paired with DNA target sites and trained inDelphi, a machine learning model that predicts genotypes and frequencies of 1- to 60–base–pair deletions and 1–base–pair insertions with high accuracy (r = 0.87) in five human and mouse cell lines. inDelphi predicts that 5–11% of Cas9 guide RNAs targeting the human genome are ‘precise–50’, yielding a single genotype comprising greater than or equal to 50% of all major editing products. We experimentally confirmed precise–50 insertions and deletions in 195 human disease–relevant alleles, including correction in primary patient–derived fibroblasts of pathogenic alleles to wild–type genotype for Hermansky–Pudlak syndrome and Menkes disease. This study establishes an approach for precise, template–free genome editing.

Clustered regularly interspaced short palindromic repeats (CRISPR)–Cas9 has revolutionized genome editing, providing powerful research tools and promising agents for the potential treatment of genetic diseases1–3. The DNA-targeting capabilities of Cas9 have been improved by the development of guide RNA (gRNA) design principles4, modeling of factors leading to off-target DNA cleavage, enhancement of Cas9 sequence fidelity by modifications to the nuclease and gRNA, and the evolution or engineering of Cas9 variants with alternative PAM sequences5. Similarly, control over the product distribution of genome editing has been advanced by the development of base editing to achieve precise and efficient single-nucleotide mutations6,7, and the improvement of template-directed homology-directed repair (HDR) of double-stranded breaks8. Despite these developments, base editing does not mediate insertions or deletions, and HDR is limited by low efficiency, particularly in non-dividing cells, and by undesired by-products. As many human genetic variants associated with disease arise from insertions and deletions9,10, methods to efficiently introduce insertions and deletions to alleviate pathogenic mutations in a predictable manner with a major single-genotype outcome would advance the field of genome editing.

Non-homologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ) processes are major pathways involved in the repair of Cas9-mediated double-stranded breaks that can result in highly heterogeneous repair outcomes comprising hundreds of repair genotypes. Although end-joining repair of Cas9-mediated double-stranded DNA breaks has been harnessed to facilitate knock-in of DNA templates1,11,12 or deletion of intervening sequence between two cleavage sites3, NHEJ and MMEJ are not generally considered useful for precision genome editing applications. Previous work has found that the heterogeneous distribution of Cas9-mediated editing products at a given target site is reproducible and dependent on local sequence context13,14, but no general methods have been described to predict genotype products following Cas9-induced double-stranded DNA breaks.

In this study, we developed a high-throughput Streptococcus pyogenes Cas9 (SpCas9)–mediated repair outcome assay to characterize end-joining repair products at Cas9-induced double-stranded breaks using 1,872 target sites based on sequence characteristics of the human genome. We used the resulting rich set of repair product data to train inDelphi, a machine learning algorithm that accurately predicts the frequencies of the substantial majority of template-free Cas9-induced insertion and deletion events at single-base resolution (https://indelphi.giffordlab.mit.edu/). We find that, in contrast to the notion that end-joining repair is heterogeneous, inDelphi identifies that 5–11% of SpCas9 gRNAs in the human genome induce a single predictable repair genotype in ≥50% of editing products. Building on this idea of precision gRNAs, we used inDelphi to design 14 gRNAs for high-precision template-free editing yielding predictable 1-bp insertion genotypes in endogenous human disease-relevant loci and experimentally confirmed highly precise editing (median 61% among edited products) in two human cell lines. We used inDelphi to reveal human pathogenic alleles that are candidates for efficient and precise template-free gain-of-function genotypic correction and achieved template-free correction of 183 pathogenic human microduplication alleles to the wild-type genotype in ≥50% of all editing products. Finally, we integrate these developments to achieve high-precision correction of five pathogenic low-density lipoprotein receptor (LDLR) microduplication alleles in human and mouse cells, as well as correction of endogenous pathogenic microduplication alleles for Hermansky–Pudlak syndrome (HPS1) and Menkes disease (ATP7A) to the wild-type sequence in primary patient-derived fibroblasts.
High-throughput assaying of Cas9-mediated DNA repair products supports the design of the inDelphi model. a. A high-throughput genome-integrated library for assaying Cas9 editing products. b. Categories of editing products at 1,996 lib-A target sites in mESCs. c. Categories of editing products at 1,996 VO endogenous target sites in HEK293 cells. d. Mechanism of microhomology-mediated end-joining repair. e. inDelphi uses machine learning to predict the frequencies of editing products from target DNA sequence (selected outcomes depicted in table). Major editing outcomes include +1 to −60 indels.

Template-free Cas9 editing is predictable

To capture Cas9-mediated end-joining repair products across a wide variety of target sites, we designed a genome-integrated gRNA and target library screen in which many unique gRNAs are paired with 53-bp target sites containing a single canonical ‘NGG’ SpCas9 protospacer-adjacent motif (PAM) that directs cleavage to the centre of each target site (Fig. 1a). Previously reported repair products at 90 loci in three human cell lines (HCT116, K562, and HEK293) (Fig. 1b, c), as well as VO endogenous target sites in HEK293 cells (Fig. 1d, e), constitute 80–95% of all observed editing products (Fig. 1b, c). Three repair classes were defined: deletions, insertions, and single-base (1-bp) insertions; we define these three repair classes as constituting all major editing outcomes. The insertion and deletion (indel) frequencies at 86 target sites were consistent between endogenous data in HEK293, K562 and HCT116 cells and lib-A data in mESCs and U2OS cells (median r = 0.65 to 0.76 for pairs of cell types when adjusting for 1-bp insertion frequencies, median r = 0.52 to 0.76 without adjustment, Extended Data Fig. 1). Together, these data confirm that Cas9-mediated editing products from our library assay reflect previously reported endogenous editing in human cells.

Using lib-A, we designed a new machine learning model, inDelphi, to predict the frequency of all major editing outcomes at any given target site. This model consists of three interconnected modules aimed at predicting MH deletions, MH-less deletions and 1-bp insertions (Fig. 1e). inDelphi predicts MH deletions using a module that simulates the MMEJ repair mechanism, in which 5′–3′ end resection at a double-stranded break reveals two 3′ single-stranded DNA overhangs that can anneal through sequence microhomology. Extraneous single-stranded DNA overhangs are eliminated, and DNA synthesis and ligation generates a double-stranded DNA repair product (Fig. 1f).

Through this mechanism, each microhomology results in a distinct deletion genotype (Fig. 1d, Supplementary Discussion). inDelphi assigns a score (phi) to a candidate microhomology based on a neural-network-learned score using its length and GC content with a penalty based on the deletion length. Relative frequencies are obtained by normalizing the phi scores of microhomologies of interest to sum to one, thereby modelling MH deletions as a competitive process.

InDelphi models deletions inconsistent with MMEJ with a second neural network module that predicts the total frequency of groups of MH-less deletion outcomes using the minimum required resection length as the only input feature (Fig. 1e). We hypothesize that MH-less deletions arise primarily from the classical and alternative NHEJ pathways (Supplementary Discussion).

The MH and MH-less neural networks were jointly trained using data from 1,095 lib-A target sites in mESCs with backpropagation in a multi-task manner to predict both deletion length frequencies and estimated precision of deletions (Supplementary Methods) that identified an average of 245 unique repair outcomes with high confidence (Supplementary Methods) per target site in mESCs (45 in U2OS cells) after adjusting with control data. Repair outcomes in experimental replicates within the same cell type were consistent (median r = 0.89 in mESCs, 0.77 in U2OS cells, Extended Data Fig. 1).
Computational experiments confirmed that the design of the two-sided highly influential and causal factor for 1-bp insertion repair. Consistent with lib-A (Fig. 2e), suggesting that local sequence context is a variation in 1-bp insertion frequency (from $\leq$ background sequences with weak microhomology revealed substantial four nucleotides surrounding the Cas9 cleavage site in three constant of deletions (Methods) were significantly more likely to yield insertions at the expense of deletions ($P < 2.0 \times 10^{-3}$, Extended Data Fig. 3). Randomization of four nucleotides surrounding the Cas9 cleavage site in three constant background sequences with weak microhomology revealed substantial variation in 1-bp insertion frequency (from $\leq$5% to $\geq$80% of all edited products, Fig. 2d, Extended Data Fig. 3) and identified mini-motifs consistent with lib-A (Fig. 2e), suggesting that local sequence context is a highly influential and causal factor for 1-bp insertion repair.

On the basis of these data, inDelphi models insertions and deletions as competitive processes in which microhomology strength and precision of deletions influence the relative frequency of 1-bp insertions, and local sequence context influences the relative frequency and genotypic outcomes of 1-bp insertions (Fig. 1c). inDelphi makes predictions within each module in a cell-type-invariant manner, only using cell-type-specific data to predict the overall ratio of 1-bp insertions to deletions. Collectively across all three modules, inDelphi predicts the indel lengths of 80–95% of Cas9-mediated editing products and the genotypes of 65–80% of all products (Fig. 3a, Extended Data Fig. 4) from sequence context alone.

InDelphi achieves high accuracy in predicting genotype frequencies (median $r = 0.94$) and indel length frequency distributions (median $r = 0.91$) in 189 held-out lib-A target sites in mESCs (Extended Data Fig. 4), with similarly high accuracy in U2OS cells (median $r = 0.88$ and 0.91, Extended Data Fig. 4). On held-out endogenous data, inDelphi also performed strongly on the two tasks (median $r = 0.87$ and 0.84 across 87–90 target sites in K562, HCT116 and HEK293 cells, Fig. 3b, c). Taken together, these results establish that inDelphi to facilitate Cas9-mediated gene knockout approaches by allowing a priori selection of gRNAs that induce high or low knockout frequencies. We note that microhomology deletions in human exons have a significant tendency to remain in-frame compared to non-coding human DNA (Extended Data Fig. 4).

**Highly precise template–free Cas9 editing**

Although end-joining repair is highly efficient in inducing mutations after Cas9 treatment, its propensity to induce a heterogeneous mixture of repair genotypes has limited applications for precision genome editing.

We used inDelphi to estimate the fraction of SpCas9 gRNAs targeting exons and introns in the human genome that support precise end-joining repair. Defining ‘precision’-x’ gRNAs as those predicted to produce a single genotypic outcome in $\geq$4% of all major editing outcomes proximal to the cleavage site, inDelphi predicts that 28% and 47% of gRNAs are precision-30, whereas 5% and 11% of gRNAs are precision-50, when trained on mESC and U2OS cell data, respectively (Fig. 3f, Extended Data Table 1).

To test the accuracy of the predictions made using inDelphi of precise repair in endogenous settings, we selected 14 SpCas9 gRNAs predicted to induce precision-40 1-bp insertions. We delivered SpCas9 with gRNAs and performed endogenous HTS in human U2OS and HEK293T cells. We observed that 10 out of 14 predicted precision-40 1-bp insertion gRNAs induced a single 1-bp insertion genotype in $\geq$40% of edited products with an overall significantly higher precision ($P < 4.2 \times 10^{-8}$) than baseline data in HEK293T (median 55% compared with 25% baseline in VO target sites in HEK293) and U2OS cells (median 57% compared with 14% baseline in lib-A, U2OS, Fig. 3e). We similarly validated 10 gRNAs for high-precision deletions with endogenous HTS in both cell types (Extended Data Table 2). Collectively, these observations establish the ability of inDelphi to identify, from sequence features alone, gRNAs that induce significantly more precise editing than the general population of gRNAs.
Fig. 3 | inDelphi accurately predicts nearly all editing outcomes. 

a, Fraction of endogenously editing products given predictions in HEK293 (n = 86 target sites), HCT116 (n = 91) and K562 cells (n = 82).

b, Predictive performance on endogenously observed frequencies of genotypes (b) and indel lengths (c) in HEK293 (median, 0.87 and 0.84), HCT116 (median, 0.87 and 0.83), and K562 (median, 0.83 and 0.79) cells. The box denotes the 25th, 50th and 75th percentiles, and whiskers show 1.5 times the interquartile range. 

c, Comparison of predictions from two methods to observed frame frequencies (n = 86 target sites, HEK293 cells), regression estimate ± 95% confidence intervals. 

d, Predicted frequency among major editing products using mESC-trained inDelphi (%). 

e, Predicted frequency among major editing products using U2OS-trained inDelphi (%). 

Template-free correction of pathogenic alleles

We used inDelphi to identify new targets for therapeutic genome editing. Starting with 23,018 pathogenic short indels (ClinVar and HGMD databases), we used inDelphi to identify pathogenic alleles that are suitable for template-free Cas9-mediated editing to effect precise gain-of-function editing of the pathogenic genotype. We pursued two genic disease allele categories that have not been previously identified as targets for Cas9-mediated repair: pathogenic frameshifts in which inDelphi predicts that 50–90% of Cas9-mediated deletion products will correct the reading frame (mean baseline frequency of 34% among disease-associated frameshift mutations) and pathogenic microduplication alleles in which a short sequence duplication leads to a frameshift mutation or disrupts protein function and which inDelphi predicts can be repaired to wild-type genotype in a large fraction of Cas9 editing products (Fig. 4a).

We selected 1,592 pathogenic human loci with high predicted rates of frame correction or microduplication correction to the wild-type sequence for inclusion in a second library (lib-B). We observed that 183 human disease microduplication alleles included in lib-B were repaired to wild-type in ≥50% of all products (Fig. 4b), and 508 pathogenic human frameshift alleles were corrected into proper reading frames in ≥50% of all products in mESCs (Fig. 4c), in agreement with inDelphi’s predictions (r = 0.64 and 0.64). We observed similar results in U2OS cells (r = 0.65 for frame correction, r = 0.61 for genotype correction to wild type, Extended Data Fig. 5). Although repair to the wild-type genotype unambiguously restores wild-type protein function, we note that frame correction that alters the coding sequence requires case-by-case analysis to validate rescue of protein function.

To determine whether the efficiency of microduplication repair can be increased by manipulation of DNA repair pathways, we performed Cas9 cleavage of lib-B in four NHEJ-deficient conditions: Prkdc−/−/Lig4−/− mESCs, and mESCs treated separately with DNA-dependent protein kinase inhibitor III (DPKi3), NU7026, and MLN4924. In NHEJ-impaired cells, the fraction of deletion outcomes not involving MH significantly decreased (median 23% to 10% with Prkdc−/−/Lig4−/−, P = 1.0 × 10−8) and 23% to 19% with DPKi3 and NU7041, P < 5.5 × 10−5 (Extended Data Fig. 6). 

In Prkdc−/−/Lig4−/− mESCs, the increased propensity towards MH deletions enabled a subset of pathogenic alleles to be repaired to wild type with markedly higher precision. Compared to wild-type mESCs in which 183 pathogenic alleles corrected to wild type in ≥50% of all edited products and 11 pathogenic alleles corrected to wild type in ≥70% of all edited products, in Prkdc−/−/Lig4−/− mESCs, 286 pathogenic alleles corrected to wild type in ≥50% of all edited products and 153 pathogenic alleles corrected to wild type in ≥70% of products (Fig. 4d, Supplementary Table 1) without increase in the rate of apoptosis (Extended Data Fig. 6). DPKi3 or NU7041 treatment also increased precise microduplication repair (Extended Data Figs. 5, 6). Taken together, impairing NHEJ can further increase the precision of wild-type correction for a large subset of pathogenic microduplications in genes such as PKD1 (corrected in 92% of edited Prkdc−/−/Lig4−/− mESCs alleles), MSH2 (88%) and LDLR (87%), supporting a model of competing end-joining repair mechanisms.

We further tested inDelphi’s prediction of highly efficient correction in a functional assay with pathogenic LDLR microduplication alleles that cause dominantly inherited familial hypercholesterolemia22. We separately introduced five pathogenic LDLR microduplication alleles within a full-length LDLR coding sequence upstream of a P2A-GFP cassette into the genome of human and mouse cells, such that Cas9-mediated repair to the wild-type LDLR sequence should induce phenotypic gain of LDL uptake and restore the reading frame of GFP. We then delivered Cas9 and a gRNA that is specific to each pathogenic allele and does not target the wild-type repaired sequence. We observed robust restoration of LDL uptake as well as restoration of GFP fluorescence in mESCs, U2OS cells and HCT116 cells in up to 79% of cells following transfection with Cas9 and inDelphi gRNAs (Fig. 4e, f, Extended Data Fig. 7). HTS confirmed efficient correction of these five LDLR microduplication alleles to wild type in human and mouse cells, as well as pathogenic microduplication alleles in the GAA, GLBI and PORCN genes introduced to cells using the same method (Extended Data Table 3). Importantly, in these experiments, we observed high-frequency LDLR phenotypic correction when cutting with either SpCas9 or StrepCas9 aureus Cas9 (SaCas9)23 (Extended Data Table 3).
This work establishes that the prediction and judicious application to accurately predict repair genotypes from other designer nucleases. Incorporating the frequencies of long deletions and translocations into predictive models of Cas9 outcomes will be an important next step. Precise template-free correction of pathogenic alleles. Effort correction of a pathogenic allele to wild type. Comparison among pathogenic alleles of observed and predicted frequencies of repair to wild-type genotype and frame. Wild-type repair frequencies of pathogenic alleles with predicted frequency ≥50% among all major editing outcomes in mESCs. Dashed lines indicate means. For mESCs containing the LDLRdup:1662_1669dupCTCTGCTGCTGA -P2A–GFP allele, flow cytometric contour plots (e) and fluorescence microscopy (f). Representative data for n = 2 independent biological replicates. Major editing outcomes include +1 to −60 indels.

Discussion

We used the Cas9-mediated end-joining repair products of thousands of target DNA loci integrated into mammalian cells to train a machine learning model, inDelphi, that accurately predicts the spectrum of cut-site proximal genotypic products resulting from double-stranded break repair at a target DNA site of interest. The ability to predict Cas9-mediated products enables new precision genome editing research applications and facilitates existing applications, such as performing efficient bi-allelic gene knockout and predicting end-joining by-products of HDR. We provide an online implementation of inDelphi to predict the spectrum of Cas9-mediated products along with predicted frameshift frequencies and precision at any target site (https://indelphi.giffordlab.mit.edu/).

The inDelphi model identifies target loci in which a substantial fraction of all repair products consist of a single genotype. Our findings suggest that 28–47% of SpCas9 gRNAs that target the human genome yield a single indel genotype in ≥30% of all major repair products (precision-30), and 5–11% yield a single indel genotype in ≥50% of all major repair products (precision-50). We show experimentally that precision template-free Cas9-mediated editing can mediate efficient gain-of-function repair at hundreds of pathogenic alleles including microduplications (Fig. 4b, e, f) in cell lines and in patient-derived primary cells (Extended Data Table 3). We note that each research or therapeutic Cas9-nuclease application may require a different level of precision depending on a variety of factors including risk/reward calculations of the gene and disease in question.

Moreover, we present evidence that suppressing NHEJ augments repair of pathogenic microduplication alleles, suggesting that temporary manipulation of DNA repair pathways could be combined with Cas9-mediated editing to favour specific editing genotypes with high precision. Genome editing currently lacks flexible strategies to correct indels in post-mitotic cells because of the limited efficiency of HDR in non-dividing cells. As MMEJ is thought to occur throughout the cell cycle, inDelphi may provide access to predictable and precise post-mitotic genome editing in a wider range of cell states. Incorporating the frequencies of long deletions and translocations into predictive models of Cas9 outcomes will be an important next step to calculate the overall precision of Cas9-nuclease editing. We anticipate that, given appropriate training data, inDelphi will also be able to accurately predict repair genotypes from other designer nucleases.

This work establishes that the prediction and judicious application of template-free Cas9 nuclease-mediated genome editing offers new capabilities for the study and potential treatment of genetic diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0686-x.

Received: 27 April 2018; Accepted: 10 September 2018; Published online 7 November 2018.

1. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819–823 (2013).
2. Mali, P. et al. RNA-guided human genome engineering via Cas9. Science 339, 823–826 (2013).
3. Jinek, M. et al. RNA-programmed genome editing in human cells. eLife 2, e00471 (2013).
4. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. 34, 184–191 (2016).
5. Adli, M. The CRISPR tool kit for genome editing and beyond. Nat. Commun. 9, 1911 (2018).
6. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature 533, 420–424 (2016).
7. Gaudelli, N. M. et al. Programmable base editing of T to C in genome DNA without DNA cleavage. Nature 551, 464–471 (2017).
8. Paquet, D. et al. Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. Nature 533, 125–129 (2016).
9. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 44, D862–D866 (2016).
10. Stenson, P. D. et al. Human Gene Mutation Database: towards a comprehensive central mutation database. J. Med. Genet. 45, 124–126 (2008).
11. Suzuki, K. et al. In vivo genome editing via CRISPR/Cas9-mediated homology-independent targeted integration. Nature 540, 144–149 (2016).
12. Nakade, S. et al. Microhomology-mediated end-joining-dependent integration of donor DNA in cells and animals using TALENs and CRISPR/Cas9. Nat. Commun. 5, 5560 (2014).
13. Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, Mdel. C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat. Biotechnol. 32, 267–273 (2014).
14. van Overbeek, M. et al. DNA repair profiling reveals nonrandom outcomes at Cas9-mediated breaks. Mol. Cell 63, 633–646 (2016).
15. Ursaki, A., Morvan, G. & Kawakami, K. Functional dissection of the Tol2 transposable element identified the minimal cis-sequence and a highly repetitive sequence in the subterminal region essential for transposition. Genetics 174, 639–649 (2006).
16. Cecchetti, R., Roncinelli, B. & Andre, A. D. Repair pathway choices and consequences at the double-strand break. Trends Cell Biol. 26, 52–64 (2016).
17. Deriano, L. & Roth, D. B. Modernizing the nonhomologous end-joining repertoire: alternative and classical NHEJ share the stage. Annu. Rev. Genet. 47, 433–455 (2013).
18. Bae, S., Kweon, J., Kim, H. S. & Kim, J.-S. Microhomology-based choice of Cas9 target sites. Nat. Methods 11, 705–706 (2014).
19. Cornu, T. I., Mussolin, C. & Cathomen, T. Refining strategies to translate genome editing to the clinic. Nat. Med. 23, 415–423 (2017).
20. Davis, A. J. & Chen, D. J. DNA double strand break repair via non-homologous end-joining. Transl. Cancer Res. 2, 130–143 (2013).
21. Arbab, M., Srinivasan, S., Hashimoto, T., Geijsen, N. & Sherwood, R. I. Cloning-free CRISPR. Stem Cell Reports 5, 908–917 (2015).
22. Bourbon, M., Alves, A. C. & Sijbrands, E. J. Low-density lipoprotein receptor mutational analysis in diagnosis of familial hypercholesterolemia. Curr. Opin. Lipidol. 28, 120–129 (2017).
23. Ran, F. A. et al. In vivo genome editing using Staphylococcus aureus Cas9. Nature 520, 186–191 (2015).
24. Oh, J. et al. Positional cloning of a gene for Hermansky–Pudlak syndrome, a disorder of cytoplasmic organelles. Nat. Genet. 14, 300–306 (1996).
25. Bioh, R. et al. DNA double-strand break resection occurs during non-homologous end joining in G1 but is distinct from resection during homologous recombination. Mol. Cell 65, 671–684 (2017).
26. Shin, H. Y. et al. CRISPR/Cas9 targeting events cause complex deletions and insertions at 17 sites in the mouse genome. Nat. Commun. 8, 15464 (2017).
27. 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).

Acknowledgements The authors thank O. Juez, R. Jodhani and C. Araneo for technical assistance and the MIT Biomicro Center, the Harvard Medical School Biopolymers Facility, and the Broad Institute Genomics Platform for sequencing. The authors acknowledge funding from an NSF Graduate Research Fellowship to M.W.S.; an NWO Rubicon Fellowship to M.A.; 1R01HG010372 (C.A.C.); DARPA HR0011-17-2-0049, NIHGM1HG009490, R01 EB022376, R35 GM118062, HHMI (D.R.L.); 1K01DK101684, the Human Frontier Science Program, NWO, Brigham Research Institute, Harvard Stem Cell Institute, and American Cancer Society (R.I.S.).

Reviewer information Nature thanks D. Durocher, R. Platt and the anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.W.S., J.Y.H. and D.K.G. contributed to the inDelphi model. M.W.S., M.A., C.A.C., D.R.L., D.K.G. and R.I.S. contributed to the editing libraries, assays and applications. M.A. and R.I.S. contributed to the library experimental protocol and performed library experiments in mESCs, DNA repair-deficient mESCs, and U2OS cells. D.W., S.J.C., O.K. and R.I.S. performed endogenous experiments in mESCs, HCT116, U2OS and HEK293T cells. M.A. performed endogenous experiments in primary patient fibroblasts. M.W.S., J.Y.H., C.A.C. and D.K.G. contributed to algorithm development and computational analysis. M.W.S., M.A., D.R.L., D.K.G. and R.I.S. contributed to writing and editing the manuscript.

Competing interests The authors declare competing interests; patent applications have been filed on this work. D.R.L. is a consultant and co-founder of Editas Medicine, Beam Therapeutics and Pairwise Plants, companies that use genome editing technologies. D.K.G. is a co-founder of Think Therapeutics, a company that uses machine learning for therapeutic development.

Additional information Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0686-x.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0686-x.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.R.L. or D.K.G. or R.I.S.
Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
In brief, the cloning process involves ordering a library of oligonucleotides pairing a gRNA protospacer with its 5′-bp target site, centre on an Ngg PAM. To insert the gRNA hairpin between the gRNA protospacer and the target site, the library undergoes an intermediate Gibson Assembly circularization step, restriction enzyme linearization and Gibson Assembly into a plasmid backbone containing a U6 promoter to facilitate gRNA expression, a hygromycin-resistance cassette and flanking Tol2 transposon sites to facilitate integration into the genome.

The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cloning.

We performed qPCR by addition of SybrGreen Dye (Thermo Fisher) to determine amplification efficiency and the remainder was grown in liquid culture in DRM medium supplemented with GlycoBlue Coprecipitant (Thermo Fisher) and reconstituted in milliQ water. We ran the PCR reaction for half of the determined number of cycles at this stage. Extension time for all PCR reactions was extended to 1 min per cycle to prevent skewing towards GC-rich sequences. The 246-bp fragment was purified using a PCR purification kit (Qiagen). Separately, the donor template for circularization was amplified with NEBNext polymerase (New England Biolabs) using primers ‘oligonucleotide library forward’ and ‘oligonucleotide library reverse’ (see below), to extend the sequences with overlaps complementary to the donor template used for circular assembly. To avoid overamplification in the library cloning process, we first performed qPCR by addition of SybrGreen Dye (Thermo Fisher) to determine the number of cycles required to complete the exponential phase of amplification. We ran the PCR reaction for half of the determined number of cycles at this stage. The amplified synthetic library and donor templates were ligated by Gibson Assembly (New England Biolabs) in a 1:3 molar ratio for 1 h at 50 °C, and unligated donor-template containing plasmid (Addgene 71485)21 previously digested with BbsI and XbaI (New England Biolabs), to facilitate gRNA expression and integration into the genome.

The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

The library cloning primers.

We performed qPCR by addition of SybrGreen Dye (Thermo Fisher) to determine amplification efficiency and the remainder was grown in liquid culture in DRM medium supplemented with GlycoBlue Coprecipitant (Thermo Fisher) and reconstituted in milliQ water. We ran the PCR reaction for half of the determined number of cycles at this stage. Extension time for all PCR reactions was extended to 1 min per cycle to prevent skewing towards GC-rich sequences. The 246-bp fragment was purified using a PCR purification kit (Qiagen). Separately, the donor template for circularization was amplified with NEBNext polymerase (New England Biolabs) using primers ‘oligonucleotide library forward’ and ‘oligonucleotide library reverse’ (see below), to extend the sequences with overlaps complementary to the donor template used for circular assembly. To avoid overamplification in the library cloning process, we first performed qPCR by addition of SybrGreen Dye (Thermo Fisher) to determine the number of cycles required to complete the exponential phase of amplification. We ran the PCR reaction for half of the determined number of cycles at this stage. The amplified synthetic library and donor templates were ligated by Gibson Assembly (New England Biolabs) in a 1:3 molar ratio for 1 h at 50 °C, and unligated donor-template containing plasmid (Addgene 71485)21 previously digested with BbsI and XbaI (New England Biolabs), to facilitate gRNA expression and integration into the genome.

The linearized fragment was further amplified with NEBNext polymerase using primers ‘plasmid insert forward’ and ‘plasmid insert reverse’ (see below) for the addition of overlaps complementary to the 5′ and 3′ regions of a Tol2 transposon containing gRNA expression plasmid (Addgene 71485)21 previously digested with BbsI and XbaI (New England Biolabs), to facilitate gRNA expression and integration into the library into the genome of mammalian cells. To avoid overamplification, we performed qPCR by addition of SybrGreen Dye (Thermo Fisher) to determine the number of cycles required to complete the exponential phase of amplification, and then ran the PCR reaction for the determined number of cycles. The 375-bp amplicon was gel-purified (Qiagen) from a 2.5% agarose gel.

The linearized fragment was further amplified with NEBNext polymerase using primers ‘plasmid insert forward’ and ‘plasmid insert reverse’ (see below) for the addition of overlaps complementary to the 5′ and 3′ regions of a Tol2 transposon containing gRNA expression plasmid (Addgene 71485)21 previously digested with BbsI and XbaI (New England Biolabs), to facilitate gRNA expression and integration into the library into the genome of mammalian cells. To avoid overamplification, we performed qPCR by addition of SybrGreen Dye (Thermo Fisher) to determine the number of cycles required to complete the exponential phase of amplification, and then ran the PCR reaction for the determined number of cycles. The 375-bp amplicon was gel-purified (Qiagen) from a 2.5% agarose gel.

The 375-bp amplicon and double-digested Tol2 transposon containing gRNA expression plasmid were ligated by Gibson Assembly (New England Biolabs) in a 3:1 ratio at 1 h at 50 °C. Assembled plasmids were purified by isopropanol precipitation with GlycoBlue Coprecipitant (Thermo Fisher) and reconstituted in milliQ water and transformed into NEB10beta (New England Biolabs) electrocompetent cells. Following recovery, a small dilution series was plated to assess transformation efficiency and the remainder was grown in liquid culture in DRM medium overnight at 37 °C. A detailed step-by-step library cloning protocol is provided in the Supplementary Methods.

The plasmid library was isolated by Midiprep plasmid purification (Qiagen). Library integrity was verified by restriction digest with Sphi (New England Biolabs) for 1 h at 37 °C, and sequence diversity was validated by HTS as follows:

Library cloning primers.

Oligonucleotide library forward: 5′-TTTTTGTGTTTCTCGTGTCGTCTGAAGAAGATGGGTGAGGCAGTCAT-3′; oligonucleotide library reverse: 5′-GTTGTAATACGACTCACTATAGGG-3′; circular donor forward: 5′-GGTTTAAGATCCTGCTGCAAACGC-3′; circular donor reverse: 5′-ATGACACCCGTAACACTGG-3′; plasmid insert forward: 5′-GTGATGCTAAAGAGAAGTGCTCTTCACTGCTCTTCAAAC-3′; plasmid insert reverse: 5′-GGTTCCTCGTCAGGATTTGTCCTG-3′.

Cloning.

We performed qPCR by addition of SybrGreen Dye (Thermo Fisher) to determine amplification efficiency and the remainder was grown in liquid culture in DRM medium overnight at 37 °C. A detailed step-by-step library cloning protocol is provided in the Supplementary Methods.

The plasmid library was isolated by Midiprep plasmid purification (Qiagen). Library integrity was verified by restriction digest with Sphi (New England Biolabs) for 1 h at 37 °C, and sequence diversity was validated by HTS as follows:

Library cloning primers.

Oligonucleotide library forward: 5′-TTTTTGTGTTTCTCGTGTCGTCTGAAGAAGATGGGTGAGGCAGTCAT-3′; oligonucleotide library reverse: 5′-GTTGTAATACGACTCACTATAGGG-3′; circular donor forward: 5′-GGTTTAAGATCCTGCTGCAAACGC-3′; circular donor reverse: 5′-ATGACACCCGTAACACTGG-3′; plasmid insert forward: 5′-GTGATGCTAAAGAGAAGTGCTCTTCACTGCTCTTCAAAC-3′; plasmid insert reverse: 5′-GGTTCCTCGTCAGGATTTGTCCTG-3′.

Cloning.

We performed qPCR by addition of SybrGreen Dye (Thermo Fisher) to determine amplification efficiency and the remainder was grown in liquid culture in DRM medium overnight at 37 °C. A detailed step-by-step library cloning protocol is provided in the Supplementary Methods.

The plasmid library was isolated by Midiprep plasmid purification (Qiagen). Library integrity was verified by restriction digest with Sphi (New England Biolabs) for 1 h at 37 °C, and sequence diversity was validated by HTS as follows:

The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Library cloning primers.

Oligonucleotide library forward: 5′-TTTTTGTGTTTCTCGTGTCGTCTGAAGAAGATGGGTGAGGCAGTCAT-3′; oligonucleotide library reverse: 5′-GTTGTAATACGACTCACTATAGGG-3′; circular donor forward: 5′-GGTTTAAGATCCTGCTGCAAACGC-3′; circular donor reverse: 5′-ATGACACCCGTAACACTGG-3′; plasmid insert forward: 5′-GTGATGCTAAAGAGAAGTGCTCTTCACTGCTCTTCAAAC-3′; plasmid insert reverse: 5′-GGTTCCTCGTCAGGATTTGTCCTG-3′.

Cloning.

We performed qPCR by addition of SybrGreen Dye (Thermo Fisher) to determine amplification efficiency and the remainder was grown in liquid culture in DRM medium overnight at 37 °C. A detailed step-by-step library cloning protocol is provided in the Supplementary Methods.

The plasmid library was isolated by Midiprep plasmid purification (Qiagen). Library integrity was verified by restriction digest with Sphi (New England Biolabs) for 1 h at 37 °C, and sequence diversity was validated by HTS as follows:

Library cloning primers.

Oligonucleotide library forward: 5′-TTTTTGTGTTTCTCGTGTCGTCTGAAGAAGATGGGTGAGGCAGTCAT-3′; oligonucleotide library reverse: 5′-GTTGTAATACGACTCACTATAGGG-3′; circular donor forward: 5′-GGTTTAAGATCCTGCTGCAAACGC-3′; circular donor reverse: 5′-ATGACACCCGTAACACTGG-3′; plasmid insert forward: 5′-GTGATGCTAAAGAGAAGTGCTCTTCACTGCTCTTCAAAC-3′; plasmid insert reverse: 5′-GGTTCCTCGTCAGGATTTGTCCTG-3′.
For SpCas9 targeting experiments, cells were transduced with a single lentivirus containing an SpCas9 and sgRNA expression cassette to target SpCas9 cleavage to either the HPS1: c.1472_1487del16 or ATP7A: c.6913_6917delCTTATT microduplication locus for use in HPS1 and Menkes Syndrome fibroblasts, respectively. The lentiviral plasmids were obtained from (LV01, Sigma-Aldrich) and lentivirus was produced by the Boston Children’s Hospital Viral Core. Fibroblasts were plated in 12-well plates at 12,500 cells cm⁻² one day before transduction. Cells were treated with 10–20 µg of virus in the presence of 8 µg ml⁻¹ Polybrene (Sigma-Aldrich) on two consecutive days and collected on day 10 after transduction.

**Apopotosis analysis.** Wild-type and Prkdc⁻/⁻/Lig4⁻/⁻ mESCs with stably integrated Lib-A were transfected with p2T-CAG-SpCas9-P2A-GFP-PuroR using Lipofectamine 3000 following standard protocols in 6-well plates with 10⁵ cells. After 24 h, cells were stained with Annexin V Alexa Fluor 568 conjugate (Thermo Fisher) according to the manufacturer’s protocols. Fluorescence was detected on a Cytoflex LX (Beckman Coulter) and analysed using FlowJo (FlowJo LLC).

Deep sequencing. Genomic DNA was collected from cells after >7 week of selection. For library samples, 16 µg gDNA was used for each sample; for individual locus samples, 2 µg gDNA was used; for plasmid library verification, 0.5 µg purified plasmid DNA was used.

For individual locus samples, the locus surrounding CRISPR–Cas9 mutation was PCR-amplified in two steps using primers >50-bp from the Cas9 target site. PCR1 was performed using the primers specified below. PCR2 was performed to add full-length Illumina sequencing adapters using the NEBNext Index Primer Sets 1 and 2 (New England Biolabs) or internally ordered primers with molony-less deletions among all deletion products in wild-type (lib-A and lib-B target sites).

PCR1 was performed using the primers specified below. PCR2 was performed to add full-length Illumina sequencing adapters using the NEBNext Index Primer Sets 1 and 2 (New England Biolabs) or internally ordered primers with molony-less deletions among all deletion products in wild-type (lib-A and lib-B target sites).

**Figure 2b,** comparison of 1-bp insertion frequencies among Cas9-edited products from 1,996 lib-A target sites. *t* = 5.4 × 10⁻⁶; **t** = 8.6 × 10⁻⁸, two-sided two-sample t-test, statistic = −13.0 and −18.4, d.f. = 777 and 1,994; Hedges’s *g* = 0.94 and 0.85, for *t* and **t**, respectively.

**Figure 2e,** comparison of the 1-bp insertion frequency at sequences in Fig. 2c with varying positions −4 and −3. Box plot as in Figure 2b. *t* = 0.03; **t** = 2.98 × 10⁻⁷, two-sided two-sample t-test, statistic = −2.2 and −6.5, d.f. = 185 and 32, Hedges’s *g* = 0.58 and 2.3, for *t* and **t**, respectively.

**Figure 3e,** comparison of 1-bp insertion frequencies among edited outcomes in U2OS cells (n = 27 observations, baseline n = 1,958 target sites, P = 4.2 × 10⁻⁸, two-sided Welch’s t-test, test statistic = 7.56, d.f. = 27.78, Hedges’s *g* = 1.47) and HEK293T cells (n = 26 observations, baseline n = 89 target sites, P = 8.1 × 10⁻¹², two-sided Welch’s t-test, test statistic = 10.40, d.f. = 34.14, Hedges’s *g* = 2.89).

Extended Data Fig. 3g, box plots displaying total deletion allele score and 1-bp insertion frequencies in mESCs for 312 4-bp target sites and 89 VO sequences. *t* = 6.1 × 10⁻⁵; two-sided two-sample t-test, test statistic = −5.94, d.f. = 399, Hedges’s *g* effect size = 0.49.

Extended Data Fig. 4d, distribution of predicted frameshift frequencies among 1–60-bp deletions for SpCas9 gRNAs targeting exons (n = 1,000,294 gRNAs; mean, 66.4%) and shuffled versions (mean, 69.3%), and introns (n = 740,759) in the human genome. Dashed lines indicate means. **t** < 10⁻³⁰, two-sided Welch’s t-test, test statistic = −145.5, d.f. = 1,506,304, Hedges’s *g* = −0.19.

Extended Data Fig. 5a, comparison of microhomology deletions among all deletions at lib-B target sites in wild-type (n = 1,909 target sites), DPKi3 (n = 1,999), MLN4924 (n = 1,995), NU7026 (n = 1,999), and Prkdc⁻/⁻/Lig4⁻/⁻ (n = 1,446). Statistical tests performed against wild-type population, Welch’s two-sided two-sample t-test. *t* = 5.6 × 10⁻⁵, test statistic = 4.0, d.f. = 3,870.8, Hedges’s *g* effect size = −0.13. **t** = 3.5 × 10⁻¹³, test statistic = 7.3, d.f. = 3,890.8, Hedges’s *g* effect size = −0.23. ***t* = 5.0 × 10⁻⁴¹, test statistic = 13.6, d.f. = 2,651.6, Hedges’s *g* effect size = −0.46.

Extended Data Fig. 5b, distribution of the frequency of each class of microhomology-less deletions among all deletion products in wild-type (lib-A and lib-B target sites, n = 3,829 target sites), DPKi3 (lib-B = 1,999), MLN4924 (lib-B, n = 1,990), NU7026 (lib-B, n = 1,992) and Prkdc⁻/⁻/Lig4⁻/⁻ (lib-A and lib-B target sites, n = 3,344). P values are compared to wild type, two-sided Welch’s t-test. Comparing among unilateral top strand joining, wild type versus Prkdc⁻/⁻/Lig4⁻/⁻ (P = 1.1 × 10⁻⁹, test statistic = 20.65, d.f. = 6,223.97, Hedges’s *g* = 0.50, versus NU7026 (P = 4.3 × 10⁻⁸, test statistic = 5.50, d.f. = 2,798.38, Hedges’s *g* = 0.18). Comparing among unilateral bottom strand joining, wild type versus Prkdc⁻/⁻/Lig4⁻/⁻ (P = 4.1 × 10⁻⁸, test statistic = 17.65, d.f. = 6,479.88, Hedges’s *g* = 0.42), versus NU7026 (P = 7.7 × 10⁻⁸, test statistic = 4.48, d.f. = 2,868.90, Hedges’s *g* = 0.50). Comparing among medial joining, wild type versus MLN4924 (P = 4.6 × 10⁻¹⁰, test statistic = 10.43, d.f. = 3,240.16, Hedges’s *g* = 0.31), versus DPKi3 (P = 4.5 × 10⁻⁵, test statistic = 9.72, d.f. = 3,231.41, Hedges’s *g* = 0.29), versus NU7026 (P = 4.6 × 10⁻²¹, test statistic = 9.49, d.f. = 3,130.82, Hedges’s *g* = 0.29).

Extended Data Fig. 7f, box plot comparing observed 1-bp insertion frequency in lib-A and 12 pathogenic alleles selected by inDelphi in mESCs (combined data from n = 2 independent biological replicates). The box denotes the 25th, 50th and 75th percentiles, whiskers show 1.5 times the interquartile range, and outliers are depicted d.f. = 11.18, Hedges’s effect size = 1.47.

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

**Code availability.** All data processing, analysis and modelling code is available at www.github.com/gifford-lab/inDelphi-dataprocessinganalysis. The inDelphi model is available online at https://indelphi.giffordlab.mit.edu/.

**Data availability.** High-throughput sequencing data have been deposited in the NCBI Sequence Read Archive database under accession codes SRP141261 and SRP141444. Processed data have been deposited under the following DOIs: https://doi.org/10.6084/m9.figshare.6830816, https://doi.org/10.6084/m9.figshare.6837959, https://doi.org/10.6084/m9.figshare.6837953, and https://doi.org/10.6084/m9.figshare.6837947.

28. Kleinstiver, B. P. et al. Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. Nat. Biotechnol. 33, 1293–1298 (2015).

29. Sherwood, R. J. et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat. Biotechnol. 32, 171–178 (2014).
Extended Data Fig. 1 | Design and cloning of a high-throughput library to assess CRISPR–Cas9-mediated editing products, yielding diverse and replicate-consistent data that is concordant with repair spectra at endogenous human genomic loci. a, Empirical distributions of various predicted and measured properties of DNA from 169,279 SpCas9 gRNA target sites in the human genome. Number of target sites per range used to design lib-A are indicated. b, Cumulative percentage of endogenous deletions in VO target sites in HEK293 (n = 89 target sites), HCT116 (n = 92) and K562 (n = 86) cells that delete up to the reported number of nucleotides (x axis). c, Schematic of the cloning process used to clone lib-A and lib-B (Methods, Supplementary Discussion, Supplementary Methods). d, Number of unique high-confidence editing outcomes (Supplementary Methods) called by simulating data subsampling in data in lib-A (n = 2,000 target sites) in mESCs (combined data from n = 3 independent biological replicates) and U2OS cells (combined data from n = 2 independent biological replicates). For ‘all’, the original non-subsampled data are presented. Each box depicts data for 2,000 target sites. Outliers are not depicted. e, Pearson’s r of genotype frequencies comparing lib-A in mESCs and U2OS cells with endogenous data in HEK293 (n = 87 target sites), HCT116 (n = 68), and K562 (n = 86) cells. Outliers are depicted as diamonds. 1-bp insertion frequency adjustment was performed at each target site by proportionally scaling them to be equal between two cell types. f, Pearson’s r of genotype frequencies at lib-A target sites, comparing two independent biological replicate experiments in mESCs (n = 1,861 target sites, median r = 0.89) and U2OS cells (n = 1,921, median r = 0.77). Outliers are depicted as diamonds. Box plots denote the 25th, 50th and 75th percentiles and whiskers show 1.5 times the interquartile range.
Extended Data Fig. 2 | Categorizing and modelling Cas9-mediated DNA repair products with manual data-analysis and automated machine learning through inDelphi. a, b, Categories of Cas9-mediated genotypic outcomes in data from endogenous contexts at VO target sites in K562 (n = 88 target sites), HCT116 (n = 92), HEK293 (n = 89) cells (collectively, a) and U2OS cells (b, n = 1,958 lib-A target sites). c, Categories and defined properties (Supplementary Methods) of all sequence alignments consistent with a Cas9-mediated 7-bp deletion. d, Hypothesized mechanisms for template-free DNA repair at Cas9-mediated DSBs based on components of the classical NHEJ, alternative NHEJ or MMEJ pathways (Supplementary Discussion). e, Function learned for modelling MH deletions (Supplementary Methods). f, Function learned for modelling MH-independent deletions (MH-less-NN) mapping deletion length to a numeric score (psi, Supplementary Methods, point plot) and with deletion length penalty normalized to sum to 1 (phi, Supplementary Methods, histogram).
Extended Data Fig. 3 | Influential role of hyperlocal sequence context features in predicting and causing 1-bp insertions. a, Frequency of 1-bp insertions in mESCs (n = 1,981 lib-A target sites) and U2OS cells (n = 1,918) with varying −4 nucleotides. b, c, Plot of 1-bp insertion frequency in mESCs (n = 1,996 lib-A target sites) and U2OS cells (n = 1,966) compared to their total phi score (b) and predicted deletion length precision score (c) with Pearson’s r. d, Comparison of 1-bp insertion frequencies among all edited products from 1,966 lib-A target sites and U2OS cells (combined data from n = 2 independent biological replicates). e, Nucleotides and their effect on the frequency of 1-bp insertions in U2OS cells. Only bases with non-zero linear regression weights in 10,000-fold iterative cross-validation are shown. Total n = 1,966 lib-A target sites. f, Insertion frequency in mESCs (n = 205) and U2OS cells (n = 217) when varying four bases by the cleavage site (positions −5 to −2 counted from the NGG-PAM at positions 0–2) contained within three target sites designed with weak microhomology. g, Microhomology strength (deletion phi score) and 1-bp insertions in mESCs for 312 ‘4-bp’ target sites and 89 VO sequences. *P = 6.1 × 10⁻⁵; two-sided two-sample t-test, test statistic = −5.94, d.f. = 399, Hedges’ g effect size = 0.49. Box plots denote the 25th, 50th and 75th percentiles, whiskers show 1.5 times the interquartile range, and outliers are depicted as diamonds.
Extended Data Fig. 4 | inDelphi predictions represent nearly all editing outcomes and are accurate at predicting the frequencies of genotypes, indel lengths, and frameshift frequencies. a, b, Pearson’s r for held-out lib-A target sites comparing inDelphi predictions with observed frequencies for genotypes (a) and indel lengths (b) in mESCs and U2OS cells. The box denotes the 25th, 50th and 75th percentiles, whiskers show 1.5 times the interquartile range. Densities were smoothed with noise but do not extend beyond the data. c, Pie chart depicting the output of Delphi for specific outcome classes at lib-A target sites in mESCs. d, e, Comparison of two methods for frameshift predictions to observed values with Pearson’s r in HCT116 cells (d, n = 91 target sites) and K562 cells (e, n = 82 target sites). The error band represents the 95% confidence intervals around the regression estimate with 1,000-fold bootstrapping. f, Distribution of predicted frameshift frequencies among 1–60-bp deletions for SpCas9 gRNAs targeting exons (n = 1,000,294 gRNAs; mean = 66.4%) and shuffled versions (mean, 69.3%), and introns (n = 740,759) in the human genome. Dashed lines indicate means. ***P < 10⁻³⁰⁰, two-sided Welch’s t-test, test statistic = −145.5, d.f. = 1,506,304, Hedges’ g = −0.19.
Extended Data Fig. 5 | Characterization of lib-B data including pathogenic microduplication repair in wild-type mESCs, wild-type U2OS cells and mESCs treated with DPKi3, NU7026 and MLN4924. 

a, Box plots of the number of unique high-confidence editing outcomes (see Supplementary Methods) called by simulating data subsampling in data at 2,000 lib-B target sites in mESCs (combined data from \(n=2\) independent technical replicates) and U2OS cells (combined data from \(n=2\) independent biological replicates). In ‘all’, the full non-subsampled data are presented (see Supplementary Table 2 for read counts). Each box depicts data for 2,000 target sites. The box denotes the 25th, 50th, and 75th percentiles and whiskers show 1.5 times the interquartile range. Outliers are not depicted.

b, Frequencies of repair to wild-type genotype at 567 ClinVar pathogenic alleles versus predicted frequencies in lib-B in human U2OS cells with Pearson’s \(r\).

c, Frequencies of repair to wild-type frame at 437 ClinVar pathogenic alleles versus predicted frequencies in lib-B in human U2OS cells with Pearson’s \(r\).

d, Frequency of pathogenic microduplication repair in wild-type mESCs (\(n=1,480\) target sites) compared to mESCs treated with MLN4924 (\(n=1,569\)), NU7041 (\(n=1,561\)) and DPKi3 (\(n=1,563\)).
Extended Data Fig. 6 | Altered distributions of Cas9-mediated genotypic products in Prkdc<sup>−/−</sup>Lig4<sup>−/−</sup> mESCs and mESCs treated with DPKi3, NU7026, and MLN4924 compared to wild-type mESCs.

a, Comparison of MH deletions among all deletions at lib-B target sites in wild-type cells (n = 1,909 target sites), cells treated with DPKi3 (n = 1,999), MLN4924 (n = 1,995) or NU7026 (n = 1,999) and Prkdc<sup>−/−</sup>Lig4<sup>−/−</sup> cells (n = 1,446). Statistical tests performed against wild-type population.

Extended Data Fig. 6 continued

|        | Prkdc<sup>−/−</sup>Lig4<sup>−/−</sup> | MLN4924 | DPKi3 | NU7026 |
|--------|-------------------------------------|---------|-------|--------|
|        |野型                             |0.09     |0.16   |0.18    |
|        |MLN4924                           |         |0.77   |0.73    |
|        |DPKi3                              |         |0.81   |
|        |NU7026                             |         |

b, Comparison of the frequency of each class of MH-less deletions among all deletion products in wild-type (lib-A and lib-B target sites), DPKi3 (lib-B, n = 1,990), MLN4924 (lib-B, n = 1,980), NU7026 (lib-B, n = 1,992) and Prkdc<sup>−/−</sup>Lig4<sup>−/−</sup> (lib-A and lib-B target sites, n = 3,344). P values are compared to wild-type, two-sided Welch’s t-test. c, Frequency of 1-bp insertions at 1,055 target sites in lib-A in Prkdc<sup>−/−</sup>Lig4<sup>−/−</sup> mESCs. d, Frequencies of deletion repair to wild-type genotype in lib-B in wild-type mESCs (n = 1,480 target sites, combined data from two technical replicates) compared to conditions, with combined data from two independent biological replicates for each of Prkdc<sup>−/−</sup>Lig4<sup>−/−</sup> (n = 1,041 target sites), MLN4924 (n = 1,569), NU7026 (n = 1,561) and DPKi3 (n = 1,563). e, Table of Pearson’s r of the change in disease correction frequency compared to wild-type at n = 791 target sites for each pair of conditions. f, g, Annexin V-568 staining flow cytometry contour plots (f) and mean ± standard deviation values (g) in wild-type and Prkdc<sup>−/−</sup>Lig4<sup>−/−</sup> lib-A mESCs following transfection with SpCas9-P2A-GFP (representative data for n = 2 experiments). Box plots denote the 25th, 50th and 75th percentiles, whiskers show 1.5 times the interquartile range, and outliers are depicted as diamonds. For detailed statistics on significance tests, see Methods.
Extended Data Fig. 7 | See next page for caption.
Extended Data Fig. 7 | Template-free Cas9-nuclease editing of human and mouse cells containing pathogenic alleles. a, b, Flow cytometric contour plots showing GFP fluorescence and LDL–Dylight550 uptake in (a) and fluorescence microscopy of (b) HCT116 cells containing the denoted LDLR alleles and treated with SaCas9 and gRNA when denoted (representative data for n = 2 experiments). c, Fluorescence microscopy of U2OS cells containing the denoted LDLR alleles and treated with SaCas9 and gRNA when denoted (representative data for n = 2 experiments). d, e, Flow cytometry gating strategy used for mESC and LDLRdup–P2A–GFP untreated (d) and treated with SpCas9 and gRNA (e). f, g, Results of 12 pathogenic 1-bp deletion alleles selected by inDelphi for high 1-bp insertion frequency (combined data from n = 2 independent biological replicates) compared to lib-A (f) and presented in a table (g). The box denotes the 25th, 50th and 75th percentiles, whiskers show 1.5 times the interquartile range, and outliers are depicted as diamonds. *P = 1.6 × 10^{-4}, two-sided Welch’s t-test. For detailed statistics, see Methods. In the table, the most frequent 1-bp insertion genotype predicted by inDelphi that does not correspond to the wild-type genotype is indicated by an asterisk. In fluorescence microscopy plots, GFP fluorescence is shown in green, LDL–Dylight550 uptake in red, and Hoechst staining nuclei in blue.
Extended Data Table 1  | Frequency of gRNAs in the human genome with denoted Cas9-mediated outcome precision

| Precision-X threshold (%) | Precise product is a deletion (% of gRNAs) | Precise product is a 1-bp insertion (% of gRNAs) | Total % of gRNAs that are precise-X | Precise product is a deletion (% of gRNAs) | Precise product is a 1-bp insertion (% of gRNAs) | Total % of gRNAs that are precise-X |
|---------------------------|------------------------------------------|-----------------------------------------------|----------------------------------|------------------------------------------|-----------------------------------------------|----------------------------------|
|                           | inDelphi trained on Lib-A data from mESCs for 1-bp ins. module |                                         |                                  | inDelphi trained on Lib-A data from U2OS cells for 1-bp ins. module |                                         |                                  |
| 10                        | 82                                       | 38                                            | 93                               | 70                                       | 78                                            | 97                               |
| 15                        | 61                                       | 23                                            | 75                               | 44                                       | 64                                            | 87                               |
| 20                        | 43                                       | 15                                            | 55                               | 27                                       | 53                                            | 72                               |
| 25                        | 30                                       | 10                                            | 39                               | 17                                       | 44                                            | 58                               |
| 30                        | 21                                       | 6.6                                           | 28                               | 11                                       | 36                                            | 46                               |
| 35                        | 15                                       | 4.4                                           | 19                               | 6.9                                      | 28                                            | 34                               |
| 40                        | 10                                       | 2.9                                           | 13                               | 4.1                                      | 21                                            | 25                               |
| 45                        | 6.5                                      | 1.9                                           | 8.4                              | 2.4                                      | 15                                            | 18                               |
| 50                        | 4.3                                      | 1.3                                           | 5.6                              | 1.4                                      | 10                                            | 12                               |
| 55                        | 2.8                                      | 0.8                                           | 3.6                              | 0.8                                      | 6.7                                            | 7.5                              |
| 60                        | 1.8                                      | 0.5                                           | 2.3                              | 0.5                                      | 4.0                                            | 4.4                              |
| 65                        | 1.1                                      | 0.3                                           | 1.5                              | 0.2                                      | 2.2                                            | 2.4                              |
| 70                        | 0.7                                      | 0.2                                           | 0.9                              | 0.1                                      | 1.1                                            | 1.2                              |
| 75                        | 0.4                                      | 0.1                                           | 0.5                              | 0.04                                     | 0.5                                            | 0.5                              |
| 80                        | 0.2                                      | 0.08                                          | 0.3                              | 0.01                                     | 0.2                                            | 0.2                              |
| 85                        | 0.08                                     | 0.04                                          | 0.1                              | 0.003                                    | 0.07                                           | 0.08                             |
| 90                        | 0.03                                     | 0.02                                          | 0.05                             | 0.0007                                   | 0.03                                           | 0.03                             |

SpCas9 gRNAs in human exons and introns in mESCs (n = 1,003,524 SpCas9 gRNAs) and U2OS cells (n = 4,498,780 SpCas9 gRNAs). Predictions were smoothed with Gaussian noise (Supplementary Methods).
Extended Data Table 2 | Endogenous repair of 24 designed high-precision gRNAs in human cell lines

| Gene, exon/chr, cutisite (hg19) | Frameshift, U2OS | Most frequent genotype, U2OS | Frameshift, HEK293T | Most frequent genotype, HEK293T |
|-------------------------------|------------------|-------------------------------|---------------------|-------------------------------|
| VEGFA exon1: 458              | 91, 87           | 36, 34*                       | 90, 90              | 43, 40*                       |
| VEGFR2 exon5: 2              | 91, 91           | 50, 53*                       | 91, 91              | 50, 24*                       |
| PCD1 exon5: 208              | 90, 90           | 20, 21*                       | 91, 90              | 29, 13*                       |
| APOB exon25: 147             | 83, 83           | 22, 21*                       | 87, 85              | 35, 18*                       |
| VEGFA exon3: 127             | 85, 89           | 27, 29*                       | 93, 91              | 55, 32*                       |
| CCR5 exon1: 1941             | 82, 81           | 20, 21*                       | 86, 84              | 43, 27*                       |
| CD274 exon2: 2271            | 85, 86           | 9, 10*                        | 84, 82              | 31, 14*                       |
| APOB exon26: 5590            | 91, 89           | 28, 25*                       | 88                  | 37*                           |
| VEGFR2 exon26: 19            | 82, 82           | 35, 33*                       | 82, 82              | 40, 24*                       |
| CXCR4 exon1: 825             | 86, 86           | 32, 33*                       | 91                  | 54*                           |
| PCSK9 exon11: 15             | 81, 78           | 28, 25†                      | 78                  | 27†                           |
| CCR5 exon1: 885              | 84, 85           | 55, 52†                       | 67                  | 46†                           |
| CCR5 exon1: 1027             | 92, 94           | 61, 60†                       | 91, 92              | 49, 58†                       |
| APOB exon26: 5573            | 93, 93           | 75, 74†                       | 93, 95              | 69, 81†                       |
| CCR5 exon1: 61               | 94, 94           | 37, 25†                       | 83, 89              | 29, 38†                       |
| CCR5 exon1: 1577             | 81, 81           | 28, 29†                       | 80, 83              | 29, 43†                       |
| APOB exon22: 100             | 89, 89           | 25, 27†                       | 91, 89              | 23, 38†                       |
| APOBEC3B exon3: 202          | 83, 84           | 50, 52†                       | 75, 88              | 51, 60†                       |
| MACCHC chr1: 45973892        | 97, 95           | 80, 77†                       | 97, 98              | 78, 85†                       |
| PROK2 chr3: 71821967         | 93, 94           | 44, 41†                       | 93, 93              | 45, 53†                       |
| IDS chrX: 148564700          | 95, 95           | 72, 74†                       | 93, 95              | 64, 80†                       |
| ECM1 chr1: 150484936         | 87, 89           | 44, 47†                       | 89, 89              | 32, 35†                       |
| KCNH2 chr7: 150644566        | 40               | 25†                           | 65, 95              | 35, 14†                       |
| LDLR chr19: 11222303         | 90, 91           | 78, 77†                       | 90, 96              | 77, 83†                       |

Data from up to two independent biological replicates are depicted.

*Deletion.
†Insertion.
‡Pathogenic 1-bp insertion allele from Clinvar or HGMD.
| Pathogenic allele | #AlleleID | LDRRLdup1 | LDRRLdup2 | LDRRLdup3 | LDRRLdup4 | LDRRLdup5 | PORCNdup | GAAdup | GAAdup | GLBDdup | HPS1dup | ATP7Adup |
|------------------|----------|-----------|-----------|-----------|-----------|-----------|-----------|---------|---------|----------|---------|----------|
| LDLRdup1 | 245617 | 245706 | 245706 | 245709 | 245715 | 246266 | 25739 | 354180 | 354180 | 98805 | ND | ND |
| LDLRdup2 | 79 | 98 | 96 | 95 | 86 | 94 | 90 | 76 | 93 | 95 | ND | ND |
| LDLRdup3 | 57 | 95 | 57 | 90 | 72 | 87 | ND | 79 | 74 | 85 | ND | ND |
| LDLRdup4 | 72 | 90 | 83 | 94 | 85 | 86 | 89 | 74 | 91 | 79 | 88 | 43 |
| LDLRdup5 | 36 | 69 | 30 | 53 | 33 | 78 | ND | ND | ND | ND | ND | ND |
| PORCNdup | ND | 67 | 39 | 25 | 15 | 65 | 48 | 76 | 59 | 42 | ND | ND |
| GAAdup | ND | ND | ND | ND | ND | 77 | ND | ND | ND | ND | ND | ND |
| GAAdup | ND | ND | ND | ND | ND | 58 | 42 | ND | 63 | 41 | ND | ND |
| GLBDdup | ND | ND | ND | ND | ND | ND | ND | 88±14* | 98 |
| gRNA sequence | TGCGA | GCAAG | TCCTG | CTGCA | TTCCC | ACTC | CTGTC | AGCTG | CTGCA | GTGCA | TGTGA | CAGCA | TTTTT |
| | AGATG | GCAAA | GTCTG | AGGAC | TCATC | TACTC | CCTGG | CAGCA | GAAGG | ACTAT | GGGGA | CAGAG | CAGCC |
| | GCTCG | ATCTG | ATTTG | AAATC | AGATT | GTCTG | CTTTT | GTGTA | TGACT | GTGGC | GGCCC | TAAAG |
| | GAGCG | ACGAG | TCCGT | TGAGG | TGCTG | TAGGC | ATCCC | CTGCA | GCAGA | ATATA | CCAGA | TAAAG |
| Cas9 Type | KKH | KKH | KKH | KKH | KKH | KKH | KKH | KKH | KKH | KKH | KKH | KKH |
| | SaCas9 | SaCas9 | SaCas9 | SaCas9 | SaCas9 | SpCas9 | SpCas9 | SaCas9 | SpCas9 | SpCas9 | SpCas9 | SpCas9 |

ND, not determined. LDLRdup1, LDLR:c.526_533dupGGCTCGGA. LDLRdup2, LDLR:c.668_681dupAGGACAAATCTGAC. LDLRdup3, LDLR:c.669_680dupGGACAAATCTGAC. LDLRdup4, LDLR:c.672_683dupGAACATCTGAC. LDLRdup5, LDLR:c.1059_1071dupCCTGGCTTTTATC. GAAdup, GAA:c.2704_2716dupCAGAAAGTTGACTG. GLBDdup, GLB1:c.1456_1466dupGTCATATAT. HPS1dup, HPS:c.1472_1487dupCCTCCCTCTGCGGGG. ATP7Adup, ATP7A:c.8913_8917dupCTTAT.

*P ≤ 0.05.
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 |
|---|---|
| 0 | 1 |

- 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 | No software was used to collect the data. |
|---|---|
| Data analysis | Custom Python code used for data analysis is available at "https://github.com/maxwshen/crispr-indelphi-dataprocessinganalysis". FlowJo (FlowJo LLC) version 10.5.0 was used for fluorescence analysis. |

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

High-throughput sequencing data have been deposited in the NCBI Sequence Read Archive database under accession codes SRP141261 and SRP141144. Processed...
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

Life sciences study design

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

Sample size

Sequencing depth was chosen to obtain a minimum average of 5,000 reads per designed library oligo based on empirical measurements (using preliminary sequencing) of the distribution of coverage over library oligos.

Data exclusions

When calculating statistics of interest on processed data, sequence contexts with less than 500 reads for the relevant subset of data were excluded. For training and testing the predictive model, sequence contexts with between-replicate replicability of MH deletions below \( r = 0.85 \) were excluded. Exclusion criteria were not pre-established.

Replication

Experimental replicates were performed for Lib-A mESC, primary fibroblast HPS1, mESC PORCN, mESC GLB1, and mESC LDLR264 which also serves as a cross-celltype replicate for HCT116 LDLR264. Technical replicates were performed for Lib-B mESC, Lib-B U2OS, and Lib-B HEK293T. All experimental replicates were successful. In addition, no-Cas9 control experiments were performed for Lib-A mESC, Lib-B mESC, primary fibroblast HPS1, Lib-B U2OS, and Lib-B HEK293T. Treatment data were adjusted using control data.

Randomization

Randomized allocation of cells to experimental groups was not relevant to our study. Between treatment and control groups, the impact of cell-to-cell variability is reduced to negligible amounts by ensuring a minimum cell population diversity of 2,000 cells per designed library oligo and an average of 5,000 reads per designed library oligo.

Blinding

Blinding was not relevant to our study. Treatment data was statistically adjusted using control data.

Reporting for specific materials, systems and methods

Materials & experimental systems

| n/a | Involved in the study |
|-----|-----------------------|
| ☑   | Unique biological materials |
| ☑   | Antibodies |
| ☑   | Eukaryotic cell lines |
| ☑   | Palaeontology |
| ☑   | Animals and other organisms |
| ☑   | Human research participants |

Methods

| n/a | Involved in the study |
|-----|-----------------------|
| ☑   | ChIP-seq |
| ☑   | Flow cytometry |
| ☑   | MRI-based neuroimaging |

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

All human cell lines (HEK293T, U2OS, HCT116) were obtained from ATCC. Mouse embryonic stem cells were originally obtained as a gift from Hynek Wichterle at Columbia University. GM14609 HPS1 fibroblasts and GM13672 Menkes syndrome fibroblasts were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research.

Authentication

Human cell lines obtained from ATCC were not authenticated because all were recently obtained from ATCC which rigorously authenticates lines. Mouse embryonic stem cells were authenticated by RNA-seq profiling to compare with previous passages. Fibroblasts obtained from Coriell were not authenticated because all were recently obtained from Coriell which rigorously authenticates lines.

Mycoplasma contamination

Cells were tested every month for mycoplasma contamination and were always negative.

Commonly misidentified lines

(See ICLAC register)

No commonly-misidentified lines were used in this study.
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
All cell lines used were treated with fluorescent LDL (LDL-Dylight550) in serum-free cell culture media 3-6 hours before harvesting. Then, cell lines were trypsinized, quenched, filtered through 0.22 um filter, and run through the flow cytometer.

Instrument
BD Biosciences FACS Symphony

Software
BD Biosciences FACSDiva

Cell population abundance
No cell sorting was performed. All populations in flow cytometric analysis >1% in frequency are presented with their frequencies in the figures.

Gating strategy
Cells were first gated via FSC-A and SSC-A to define the main population (~90% of the population). Cells were next gated via SSC-A and SSC-W to define single cells (~90% of the population). Plots in figures show all cells that pass these gates.

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