LETTER

doi:10.1038/nature13695

Saturation editing of genomic regions by multiplex homology-directed repair

Gregory M. Findlay[1]*, Evan A. Boyle[1]*, Ronald J. Hause[1], Jason C. Klein[1] & Jay Shendure[1]

Saturation mutagenesis[1,2]—coupled to an appropriate biological assay—represents a fundamental means of achieving a high-resolution understanding of regulatory[3] and protein-coding[4] nucleic acid sequences of interest. However, mutagenized sequences introduced *in trans* on episomes or via random or "safe-harbour" integration fail to capture the native context of the endogenous chromosomal locus[5]. This shortcoming markedly limits the interpretability of the resulting measurements of mutational impact. Here, we couple CRISPR/Cas9 RNA-guided cleavage[6] with multiplex homology-directed repair using a complex library of donor templates to demonstrate saturation editing of genomic regions. In exon 18 of *BRCA1*, we replace a six-base-pair (bp) genomic region with all possible hexamers, or the full exon with all possible single nucleotide variants (SNVs), and measure strong effects on transcript abundance attributable to nonsense-mediated decay and exonic splicing elements. We similarly perform saturation genome editing of a well-conserved coding region of an essential gene, *DBR1*, and measure relative effects on growth that correlate with functional impact. Measurement of the functional consequences of large numbers of mutations with saturation genome editing will potentially facilitate high-resolution functional dissection of both *cis*-regulatory elements and *trans*-acting factors, as well as the interpretation of variants of uncertain significance observed in clinical sequencing.

Functional consequences of genetic variants are best studied by manipulating the endogenous locus, which provides the native chromosomal context with respect to DNA sequence and epigenetic milieu, and for proteins, endogenous levels and patterns of expression[7]. Programmable endonucleases, for example, zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs) or clustered regularly interspaced short palindromic repeat (CRISPR)/Cas-based RNA-guided DNA endonucleases, enable direct genome editing with increasing practicality[8]. However, genome editing has primarily been applied to introduce single changes to one or a few genomic loci[9], rather than many programmed changes to a single genomic locus.

We sought to leverage CRISPR/Cas9[6,10,11] to introduce saturating sets of programmed edits to a specific locus via multiplex homology-directed repair (HDR). We first targeted six bases of a *BRCA1* exon[12]. We cloned an HDR library containing random hexamers substituted at positions +5 to +10 of *BRCA1* exon 18 and fixed, nonsynonymous changes at positions +17 to +23 (as a 'handle' for selective PCR and to prevent re-cutting[13] by destroying the protospacer adjacent motif (PAM)) (Fig. 1a; Supplementary Table 1). We co-transfected pCas9-sgBRCA1x18 and the HDR library into ~800,000 HEK293T cells, achieving 3.33% HDR efficiency. We performed two independent transfections with the same HDR library ('biological replicates' 1, 2), and cells were split on day 3 ('D3 replicates' a, b).

We prepared genomic DNA (gDNA) and complementary DNA (cDNA) from bulk cells on D5. PCR reactions were primed on the 'handle' uniquely present within successfully edited genomes. Amplification was observed in HDR library/pCas9-sgBRCA1x18-transfected samples, but not in HDR library-only controls. Amplicons derived from gDNA and cDNA were deeply sequenced (Fig. 1a). The relative abundances of hexamers within replicates and the correlation between the HDR library and edited gDNA were consistent with limited 'bottlenecking' during transfection and minimal influence of hexamer identity on HDR efficiency (Extended Data Figs 1 and 2).

We estimated the effect of introducing each hexamer to these genomic coordinates on transcript abundance by calculating enrichment scores (cDNA divided by gDNA counts, calibrated to wild type). These enrichment scores were well correlated between biological replicates (Fig. 1b, 1a vs 2a: $R = 0.659$) and between D3 replicates (Extended Data Fig. 2c; 1a vs 1b: $R = 0.662$). When we pooled read counts from D3 replicates, correlation between biological replicates improved (Extended Data Fig. 2d; 1 vs 2: $R = 0.706$).

To maximize precision (see Supplementary Note 1 for discussion of reproducibility), we merged data across all four replicates for 4,048 hexamers (Fig. 1c; Supplementary Table 2). Several results support the biological validity of the resulting enrichment scores. First, as anticipated by nonsense-mediated decay (NMD), hexamers introducing stop codons were associated with markedly reduced mRNA levels (Fig. 1c; Wilcoxon rank sum test (WRST) $P = 9.7 \times 10^{-84}$; median for nonsense hexamers 12-fold below overall median). Second, previous studies measured hexamer influence on splicing at analogous coordinates of different exons via a plasmid minigene assay[14]. Despite these contextual differences, the strongest exonic splicing silencers (ESSs) (bottom 2% in ref. 14) scored ninefold below median (Fig. 1c; WRST $P = 2.0 \times 10^{-24}$), the strongest exonic splicing enhancers (ESEs) (top 2% in ref. 14) scored 1.5-fold above median (Fig. 1c; WRST $P = 2.4 \times 10^{-11}$), and the complete data sets correlated reasonably well (Extended Data Fig. 3a; $\rho = 0.524$). We also observed correlation between G+C content and enrichment scores (Extended Data Fig. 3b), strongest for bases most proximal to the splice junction, consistent with a posited role for G+C content in the stability of splicing structures[15] (although reverse transcription bias is a potential confounder).

We next sought to assay the effects of SNVs across the full 78 base pairs of *BRCA1* exon 18 (Extended Data Fig. 4). We cloned three HDR libraries with selective PCR sites in either the 5′ or 3′ region and 3% doping[16] (97(WT):1:1:1) in the other half of the exon (L: 5′ degeneracy, 3′ nonsynonymous selective PCR site; R: 3′ degeneracy, 5′ nonsynonymous selective PCR site; R2: 3′ degeneracy, 5′ synonymous selective PCR site) (Supplementary Table 1). Five days post-transfection with pCas9-sgBRCA1x18 (1.02–1.29% HDR efficiency), we selectively amplified and deeply sequenced gDNA and cDNA.

Using data from all edited exons with ≥ 1 mutation and ≥ 10 gDNA counts, we estimated effect sizes of all possible SNVs using a weighted linear model. Estimated effect sizes were reproducible ($R = 0.846$ (R), 0.853 (R2), and 0.686 (L); Fig. 2a, Extended Data Figs 5 and 6, Supplementary Table 3). Effect sizes for the same SNVs interrogated with different selective PCR strategies (R vs R2) were also well correlated ($R = 0.847$; Fig. 2b).

The estimated effect sizes reflect empirically measured changes in transcript abundance resulting from programmed edits (Fig. 2c). As expected with NMD, nonsense mutations reduced transcript abundance

[1]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA.
*These authors contributed equally to this work.

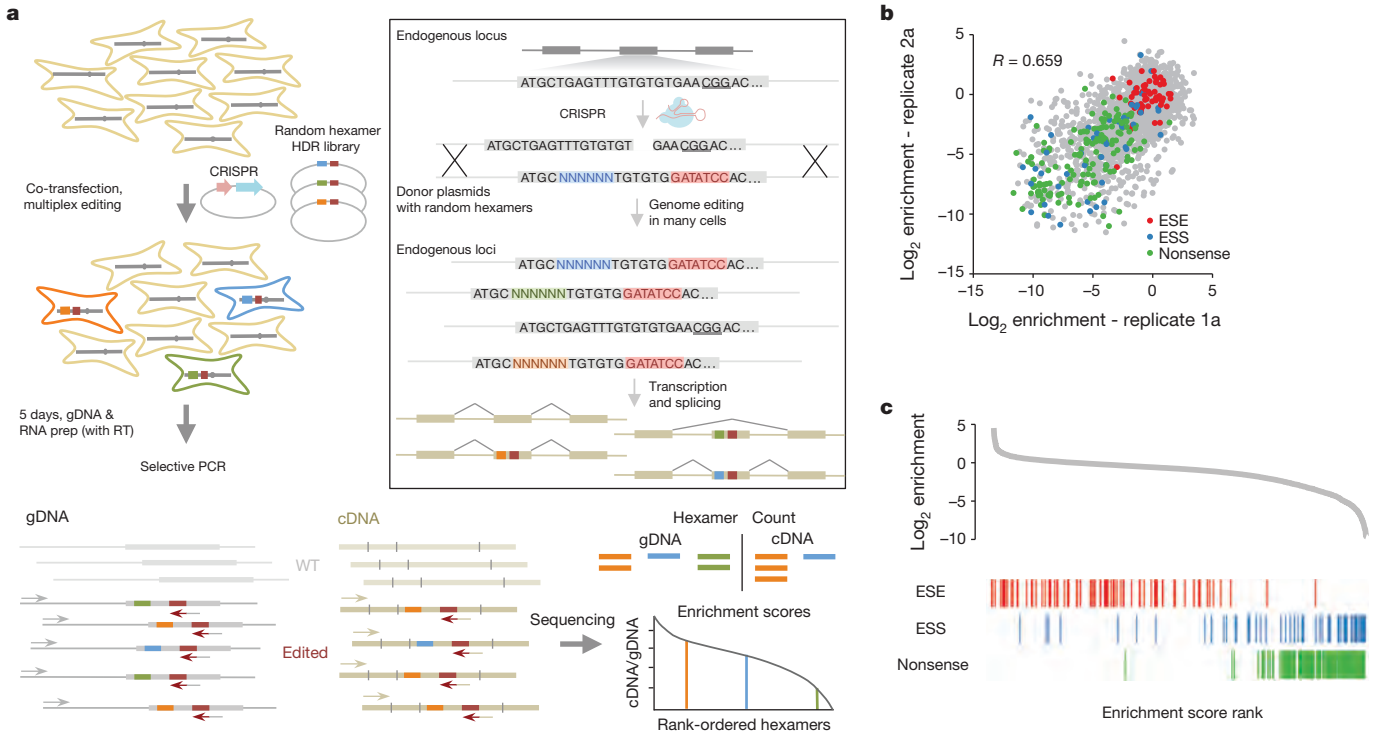


Figure 1 | Saturation genome editing and multiplex functional analysis of a hexamer region influencing *BRCA1* splicing. **a**, Experimental schematic. Cultured cells were co-transfected with a single Cas9-sgRNA construct (CRISPR) and a complex homology-directed repair (HDR) library containing an edited exon that harbours a random hexamer (blue, green, orange) and a fixed selective PCR site (red). CRISPR-induced cutting stimulated homologous recombination with the HDR library, inserting mutant exons into the genomes of many cells. At five days post-transfection, cells were harvested for gDNA and RNA. After reverse transcription, selective PCR was performed followed by sequencing of gDNA- and cDNA-derived amplicons. Hexamer enrichment scores were calculated by dividing cDNA counts by gDNA counts. **b**, Correlation of enrichment scores between biological replicates for hexamers observed in each experiment with positions of previously identified[14] exonic splicing enhancers (ESEs), exonic splicing silencers (ESSs) and stop codons indicated. **c**, Rank-ordered plot of enrichment scores with positions of ESEs, ESSs and stop codons indicated.

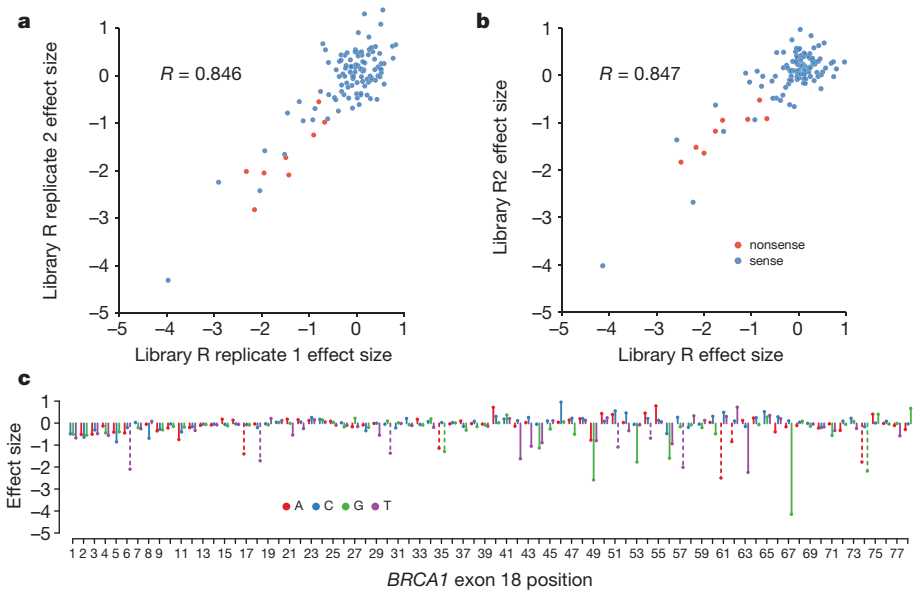


Figure 2 | Multiplex homology-directed repair reveals effects of single nucleotide variants on transcript abundance. Three separate HDR libraries (R, R2, and L) containing a 3% mutation rate (97% WT, 1% each non-WT base) in either half of *BRCA1* exon 18 were introduced to the genome via co-transfection with pCas9-sgBRCA1x18. Enrichment scores were calculated for each haplotype observed at least 10 times in gDNA sequencing, and effect sizes of SNVs were determined by weighted linear regression modelling. 'Sense' includes both missense and synonymous SNVs. **a**, Effect sizes calculated from replicate transfections of HDR library R, consisting of a 3% per-nucleotide mutation rate in the 3′-most 39 bases and the same selective PCR site used in Fig. 1, were highly correlated ($R = 0.846$). **b**, Library R2 harboured a selective PCR site composed of 5 synonymous changes, none of which are present in library R. When effect sizes derived from experiments with library R2 were plotted against those from library R, there was a strong correlation ($R = 0.847$), indicating reproducibility and demonstrating that differences between selective PCR sites did not strongly influence scores. **c**, Effect sizes for SNVs across the exon are displayed. Data sets from libraries R and L were combined to span the entire exon. Dashed lines represent SNVs that introduce nonsense codons.

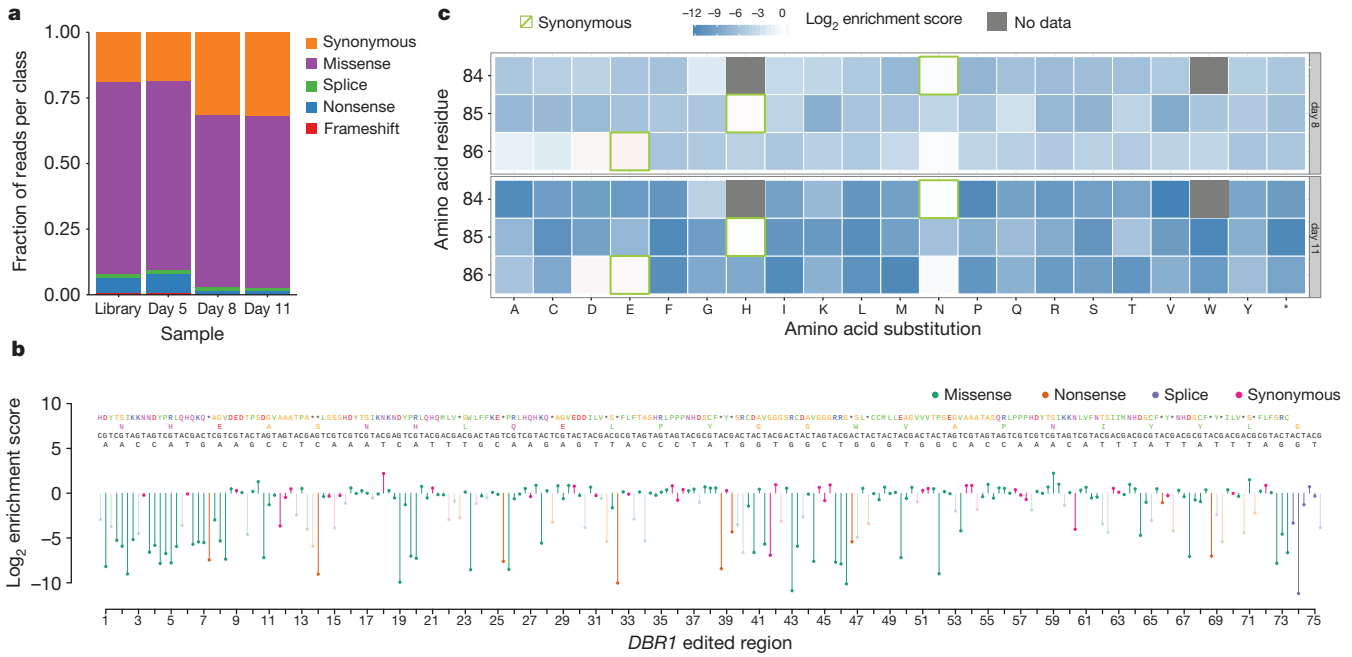


Figure 3 | Saturation genome editing and multiplex functional analysis at an essential gene, *DBR1*, in Hap1 cells. An HDR library targeting a highly conserved region of *DBR1* exon 2 was used with pCas9-EGFP-sgDbr1x2 to introduce point mutations across 75 bp and all possible codon substitutions at three residues believed to participate at the enzyme's active site. **a**, Sequencing of gDNA from the HDR library and populations of edited cells at D5, D8 and D11 reveals selection for synonymous mutations, and depletion of frameshift, nonsense and missense variants. **b**, Mean D11 enrichment scores are plotted as line segments for SNVs in the 3′-most 73 bases of exon 2 and two bases of intron 2. Above the enrichment scores in ascending order are the WT nucleotide at each position, each 1-bp genome edit, the wild-type amino acid (AA), and the AA derived from each genome edit (asterisk indicates a stop codon). Segment colour indicates mutation type, faded segments indicate discordant effects between replicates, and AAs are coloured according to the Lesk colour scheme (orange, small nonpolar; green, hydrophobic; magenta, polar; red, negatively-charged; blue, positively charged). The first nine bases shown correspond to the active site residues. **c**, D8 (top) and D11 (bottom) amino acid level enrichment scores were calculated for active site residues N84, H85, E86 after excluding discordant observations between replicates (Extended Data Fig. 10c). On both D8 and D11 we observe strong selective effects and tolerance of only synonymous (green boxes) and a few missense variants.

(WRST $P = 1.4 \times 10^{-203}$; 5.6-fold below median). Additionally, several missense and synonymous SNVs reproducibly resulted in large reductions in transcript abundance, and SNV effect sizes correlated with a predictive model for exonic variants that disrupt splicing[17] ($\rho = 0.322$; Extended Data Fig. 7a). Because HDR with library L does not destroy the PAM, we calculated enrichment scores for indels from non-homologous end-joining (NHEJ). As expected with NMD, only frameshifting indels were associated with large depletions (Extended Data Fig. 7b, c).

To further demonstrate this method, we targeted a well-conserved region of *DBR1*, the RNA lariat debranching enzyme, which scored highly in a genome-wide screen for essentiality[18] (Extended Data Fig. 8). We used array-synthesized oligonucleotides to program a *DBR1* HDR library to include the wild-type sequence and every possible SNV across 75 bp (73 3′-most bases of exon 2 and first two bases of intron 2), and also all 63 possible codon substitutions at three residues (388 genome edits were programmed; single base deletions were abundant from synthesis errors). The HDR library also introduced two fixed synonymous changes (to disrupt the PAM and prevent re-cutting[13]) and a selective PCR site in intron 2.

An optimized single-guide RNA (sgRNA) sequence[19,20] was cloned into a bicistronic sgRNA/Cas9-2A-EGFP vector (pCas9-EGFP-sgDbr1x2). Five million haploid human cells[21] (Hap1) were co-transfected with the *DBR1* HDR library and pCas9-EGFP-sgDbr1x2. On D2, ~250,000 enhanced green fluorescent protein-positive ($EGFP^+$) cells were sorted by fluorescence-activated cell sorting (FACS) and further cultured, taking samples on D5, D8 and D11 (1.14% HDR efficiency, estimated on D8). Following gDNA isolation and selective PCR, deep sequencing was performed to quantify the relative abundance of edited haplotypes in each sample.

We first examined the relative proportions of mutation classes at each time point (Fig. 3a). The strong enrichment of synonymous mutations and depletion of nonsense and frameshifting mutations over time indicated that selection was acting on edited cells in culture, consistent with *DBR1* essentiality. We calculated enrichment scores (D8 or D11 counts divided by D5 counts) for 365 of the 388 (94%) programmed edits and 12 single base deletions (the subset with relative abundance $>5 \times 10^{-5}$ on D5) (Fig. 3b; Extended Data Fig. 9; Supplementary Table 4). Enrichment scores strongly correlated with functional consequence. The median enrichment score for synonymous edits was nearly identical to wild-type (1.006-fold lower), but 73-fold lower for missense edits ($P = 1.7 \times 10^{-8}$; WRST against synonymous edits), 207-fold lower for nonsense edits ($P = 1.9 \times 10^{-9}$), and 211-fold lower for frameshifting single base deletion edits ($P = 1.5 \times 10^{-8}$). Furthermore, enrichment scores for SNVs were inversely correlated with metrics of predicted deleteriousness like CADD[22] ($\rho = -0.295$; $P = 1.2 \times 10^{-5}$; Extended Data Fig. 10a, b). Residues N84, H85 and E86 of *DBR1* were edited to all 63 possible non-wild-type codons. Consistent with their predicted role in the active site of an essential enzyme[23], only synonymous mutations and a few missense substitutions were tolerated (Fig. 3c).

Amino acid level enrichment scores were well correlated between D11 biological replicates ($R = 0.752$; $P = 2.6 \times 10^{-40}$; Extended Data Fig. 10c), and were bimodally distributed in each replicate, allowing broad classification of changes as tolerated or deleterious. The small proportion of discordantly classified variants might be explained by Hap1 reversion to diploidy or off-target effects, highlighting the importance of biological replicates for this experimental design (Supplementary Note 1). Notably, there were no reproducibly tolerated nonsense or frameshifting edits. Overall, these data support the conclusion that our empirically derived enrichment scores reflect true biological effects of specific genomic point mutations within *DBR1*.

We demonstrate that it is feasible to generate and functionally analyse hundreds to thousands of programmed genome edits at a single

locus in a single experiment. We emphasize three major limitations of the method as it stands. First, we only introduced programmed edits to the immediate vicinity of coordinates targeted by the endonuclease (Extended Data Figs 5a and 9a), and the narrow window associated with HDR mechanisms in mammalian cells[24] may fundamentally limit the size of the region that can be subjected to multiplex editing in one experiment. Saturation genome editing of a full gene—for example, to measure functional consequences of all possible variants of uncertain significance—will require multiple experiments tiling along its exons.

Second, only a small proportion of cells were successfully edited in each experiment, bottlenecking complexity, limiting reproducibility (Supplementary Note 1), and necessitating the selective PCR site. Looking forward, a variety of techniques, for example, transient hypothermia[25] or oligonucleotide-based HDR[26], may improve editing efficiency. Consistent with this, we note that ZFNs and TALENs have demonstrated efficiencies up to 50% in some studies[27,28]. Also, although the low editing efficiency necessitated using haploid cells for *DBR1* mutagenesis, this could potentially have been performed in diploid cells by knocking out one allele via NHEJ and then knocking in the HDR library to the other allele.

Finally, the development of functional assays that are biologically relevant and technically viable remains a challenge. Here, we exploited strategies that directly linked genotype to phenotype—for example, targeted RNA sequencing to measure transcript abundance or targeted DNA sequencing to measure reduced cellular fitness. Analogous approaches can be taken in other contexts—for example, targeted chromatin immunoprecipitation-sequencing (ChIP-seq) of co-activators to assay enhancers, increased cellular growth rate to assay cancer drivers or drug resistance[29], or FACS-based phenotypic sorting for cellular assays more generally[30] (Supplementary Note 2).

There is a strong demand for techniques that accurately and scalably measure mutational consequences, and a dearth of experimental data measuring distributions of effect sizes or corresponding to direct manipulation of the genome. By multiplexing both the introduction and assaying of mutations in their native context, we anticipate that saturation genome editing will accelerate our ability to measure and interpret the functional consequences of genetic variation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 May; accepted 18 July 2014.
Published online 20 August 2014.

1. Myers, R. M., Tilly, K. & Maniatis, T. Fine structure genetic analysis of a beta-globin promoter. *Science* **232,** 613–618 (1986).
2. Cunningham, B. C. & Wells, J. A. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* **244,** 1081–1085 (1989).
3. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nature Biotechnol.* **27,** 1173–1175 (2009).
4. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nature Methods* **7,** 741–746 (2010).
5. Botstein, D. & Shortle, D. Strategies and applications of *in vitro* mutagenesis. *Science* **229,** 1193–1201 (1985).
6. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).
7. Gibson, T. J., Seiler, M. & Veitia, R. A. The transience of transient overexpression. *Nature Methods* **10,** 715–721 (2013).
8. Gaj, T., Gersbach, C. A. & Barbas, C. F. III. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31,** 397–405 (2013).
9. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153,** 910–918 (2013).
10. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339,** 823–826 (2013).
11. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823 (2013).
12. Mazoyer, S. *et al.* A BRCA1 nonsense mutation causes exon skipping. *Am. J. Hum. Genet.* **62,** 713–715 (1998).
13. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnol.* **32,** 347–355 (2014).
14. Ke, S. *et al.* Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.* **21,** 1360–1374 (2011).
15. Zhang, J., Kuo, C. C. & Chen, L. GC content around splice sites affects splicing through pre-mRNA secondary structures. *BMC Genomics* **12,** 90 (2011).
16. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers *in vivo*. *Nature Biotechnol.* **30,** 265–270 (2012).
17. Mort, M. *et al.* MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biol.* **15,** R19 (2014).
18. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343,** 80–84 (2014).
19. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnol.* **31,** 827–832 (2013).
20. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8,** 2281–2308 (2013).
21. Carette, J. E. *et al.* Haploid genetic screens in human cells identify host factors used by pathogens. *Science* **326,** 1231–1235 (2009).
22. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genet.* **46,** 310–315 (2014).
23. Khalid, M. F., Damha, M. J., Shuman, S. & Schwer, B. Structure-function analysis of yeast RNA debranching enzyme (Dbr1), a manganese-dependent phosphodiesterase. *Nucleic Acids Res.* **33,** 6349–6360 (2005).
24. Elliott, B., Richardson, C., Winderbaum, J., Nickoloff, J. A. & Jasin, M. Gene conversion tracts from double-strand break repair in mammalian cells. *Mol. Cell. Biol.* **18,** 93–101 (1998).
25. Doyon, Y. *et al.* Transient cold shock enhances zinc-finger nuclease-mediated gene disruption. *Nature Methods* **7,** 459–460 (2010).
26. Chen, F. *et al.* High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nature Methods* **8,** 753–755 (2011).
27. Reyon, D. *et al.* FLASH assembly of TALENs for high-throughput genome editing. *Nature Biotechnol.* **30,** 460–465 (2012).
28. Carroll, D. Genome engineering with targetable nucleases. *Annu. Rev. Biochem.* **83,** 409–439 (2014).
29. Smurnyy, Y. *et al.* DNA sequencing and CRISPR-Cas9 gene editing for target validation in mammalian cells. *Nature Chem. Biol.* **10,** 623–625 (2014).
30. Kinney, J. B., Murugan, A., Callan, C. G., Jr & Cox, E. C. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl Acad. Sci. USA* **107,** 9158–9163 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank F. Zhang and his laboratory for the CRISPR/Cas9 backbone constructs used in this study and G. Church and his laboratory for providing reagents used to establish CRISPR/Cas9 editing techniques in our lab. We also thank members of the Shendure laboratory for helpful discussions and D. Prunkard for assistance with FACS. This work was supported by the National Institutes of Health (DP1HG007811 to J.S.) and the UW Medical Scientist Training Program (G.M.F. and J.K.).

Author Contributions The project was conceived and designed by G.M.F. and J.S. G.M.F. and E.A.B. performed experiments. E.A.B. and R.J.H. performed data analysis and generated data figures. G.M.F. generated schematic figures. G.M.F., E.A.B., R.J.H. and J.S. wrote the manuscript. J.C.K. assisted G.M.F to establish genome editing techniques in the laboratory.

Author Information Sequence data used for this analysis are available in SRA under accession number SRP044126. Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.S. (shendure@uw.edu) or G.M.F. (gf2@uw.edu).

METHODS

***BRCA1* experimental design.** As a proof-of-principle experiment, we chose to target an exon in a clinically relevant gene in which known mutations cause aberrant splicing. Previous molecular studies of a G to T nonsense mutation occurring naturally in cancer patients at chr17:41215963 suggested exon skipping[12] was secondary to the creation of an exonic splicing silencer site[31]. From this, we hypothesized that saturation genome editing of this exon could result in a wide range of splicing outcomes.

A chief consideration when performing parallel functional analysis of complex allelic series is the challenge of associating each of many mutations with the biological effects they produce. This task is more difficult when attempting such approaches at the endogenous genomic locus with limited editing efficiencies. By performing these experiments in an exon and focusing on the effects of mutations on transcript abundance, we directly link genotype and phenotype by observing the frequency of each genome edit in the transcript pool relative to its frequency in genomic DNA. This design is advantageous because it requires no specialized (that is, gene-specific) functional assay, thus making it amenable to interrogation of transcribed variants' effects on splicing/transcript abundance in any gene.

Rationale for including selective PCR sites. Given the modest proportion of HDR-edited loci in a given experiment and the high number of variants that we set out to interrogate (that is, hundreds to thousands), it would require a large amount of sequencing to sufficiently sample every variant in gDNA and cDNA pools from a population of cells that are predominantly unedited or harbouring products of NHEJ. Furthermore, at such efficiencies, the rate of error in high-throughput sequencing is high enough to obscure signal from single nucleotide variants (SNVs) (unpublished observations). Therefore, until better methods exist to isolate populations of cells successfully edited with HDR, techniques to selectively sequence molecules derived from edited cells are likely to be advantageous.

To implement this, we designed our HDR libraries to include short, fixed edits to serve as unique priming sites in genomes that successfully undergo HDR. PCR reactions primed at this site, therefore, should only amplify material from edited cells, thus reducing both the noise associated with error from sequencing unedited material and the cost of sequencing in each experiment. Additionally, we predicted that selective PCR sites that mutate the PAM and protospacer sequences would prevent Cas9 from re-cutting HDR-edited genomes. This should have the effect of increasing the proportion of cells bearing experimentally informative edits, and given the bottleneck imposed by limitations on how many successfully edited cells can be sampled, should result in more robust experimental signal.

***DBR1* experimental design.** To demonstrate that saturation genome editing can be used to explore effects of mutations on protein function and cellular fitness, we targeted *DBR1*, a well-conserved gene that scored highly in a human haploid cell genome-wide loss-of-function screen for essentiality[18]. Using haploid cells prevents gene compensation from an unedited copy[21]. Not knowing how sensitive the cells would be to mutations, we chose to target a region of exon 2 that was highly conserved, included in all transcript annotations on the UCSC Genome Browser, and coded for at least two residues (N84, H85) predicted to participate at the enzyme's active site[23]. Selection against edited cells in culture allows phenotype to be linked to genotype from sequencing of the gDNA pool over a series of time points. During HDR library construction, we designed a selective PCR site in a downstream intron to minimize any effect on gene function, and used two synonymous mutations to abrogate Cas9 re-cutting.

Given the lower transfection efficiency of Hap1 cells (~4% for the plasmids used here), we cloned a *DBR1*-targeting CRISPR construct that expressed EGFP with Cas9 and used FACS to sort a population of successfully transfected cells. The sgRNA was designed using the Zhang laboratory tool (http://crispr.mit.edu/) and selected to minimize off-target effects that could potentially impair cellular fitness[19].

HDR library and Cas9-sgRNA cloning. A homology-directed repair (HDR) library containing all possible 4,096 DNA hexamers substituted at positions +5 to +10 of *BRCA1* exon 18 (chr17:41215962–41215967; CCDS11453.1) was constructed using a partially degenerate oligonucleotide (IDT DNA; 'BRCA1ex18NNNNNN5_10selPCR') containing a 7 bp selective PCR site / EcoRV restriction digest site at position +17 to +23 (Fig. 1a, Supplementary Table 1). The oligonucleotide was PCR amplified and cloned via the In-Fusion reaction (Clontech) into a PCR-linearized pUC19-BRCA1ex18 vector containing a pre-inserted 1,573 bp fragment amplified from the surrounding BRCA1ex18 locus in HEK293T cells (chr17: 41215127–41216699) to serve as homologous arms. Additional libraries from a second degenerate oligonucleotide that was synthesized with a 3% mutation rate (97% WT, 1% each non-WT base) across the 78-bp exon were cloned similarly, such that one end of the exon would be fixed and contain either missense (as above) or synonymous mutations for selective PCR. Complete oligonucleotide and HDR library exon sequences are listed in Supplementary Table 1. All PCR reactions were performed with the KAPA HiFi HotStart ReadyMix PCR Kit.

The *DBR1* HDR library was cloned as above except with the following differences. HDR library variants were derived from 388 oligonucleotides synthesized on a microarray (CustomArray) to include all possible single base pair changes in a 75-bp region comprising part of *DBR1* exon 2 (chr3:137892342–137892416), all codon variants at the first three residues of the 75-bp region (chr3:137892408–137892416), and the reference 75-bp sequence. All *DBR1* HDR library sequences also included two synonymous mutations designed to prevent re-cutting of edited genomes by disrupting PAM and protospacer sequences (chr3:137892424 and chr3:137892421), and a 6-bp selective PCR site in intron 2 of *DBR1* (chr3:137892331–137892336). The library was cloned into a pUC19-DBR1ex2 backbone, a vector containing the surrounding *DBR1* sequence cloned from Hap1 gDNA (chr3:137891573–137893293).

A bicistronic Cas9-sgRNA vector designed to cleave within *BRCA1* exon 18 ('pCas9-sgBRCA1x18') was cloned according to a published protocol[20] by ligating annealed oligonucleotides into a human codon-optimized *Streptococcus pyogenes* Cas9-sgRNA vector from the lab of Feng Zhang (pX330-U6-Chimeric_BB-CBh-hSpCas9; Addgene plasmid #42230). The same protocol was followed to create pCas9-EGFP-sgDbr1x2 from a similar Zhang lab vector that allows for fluorescent identification of Cas9-expressing cells (pSpCas9(BB)-2A-GFP (pX458); Addgene plasmid #48138).

Cell culture and transfection. For *BRCA1* experiments, HEK293T cells were cultured in Dulbecco's modified Eagle medium (Life Technologies) supplemented with 10% FBS (AATC) and 100 U ml^{-1} penicillin + 100 μg ml^{-1} streptomycin (Life Technologies). One day before transfection, cells were split to ~40% confluency in 12-well plates with antibiotic-free media. The next day, 0.5–1.0 μg of each library was co-transfected (Lipofectamine 2000, Invitrogen) with an equivalent amount of pCas9-sgBRCA1x18. Cells were expanded to 6-well plates, then split 1:4 on day 3 into two pools, and DNA and RNA were harvested on D5 (AllPrep DNA/RNA Mini Kit, Qiagen). Biological replicates of each transfection and negative control transfections of each library without pCas9-sgBRCA1x18 were also performed.

For the *DBR1* experiment, Hap1 cells (Haplogen) were cultured in Iscove's modified Dulbecco's medium supplemented with 10% FBS and 100 U ml^{-1} penicillin + 100 μg ml^{-1} streptomycin. ~3×10^6 Hap1 cells were passaged to a 60-mm dish in antibiotic-free media one day before co-transfection with 3 μg each of pCas9-EGFP-sgDbr1X2 and the *DBR1* HDR library via Turbofectin 8.0 (OriGene) according to protocol. On D2, FACS was performed (BD FACSAria III) to isolate ~250,000 $EGFP^+$ cells which were then expanded in culture with samples taken of ~1×10^6 cells on D5, and 4–8 $\times 10^6$ cells on D8 and D11. gDNA was isolated according to protocol with the QiaAmp Kit (Qiagen). A biological replicate was performed, as well as negative controls in which the HDR library was transfected with the empty pSpCas9(BB)-2A-GFP construct (to enable FACS of transfected cells without editing).

Reverse transcription, selective PCR and sequencing. For *BRCA1* experiments, reverse transcription (RT) was performed using SuperScriptIII (Invitrogen) with a gene-specific primer located in either *BRCA1* exon 19 (hexamer experiments) or exon 21 (whole exon experiments). Initial rounds of PCR were performed on large quantities of sample gDNA (8–12 μg gDNA, 100–150 ng per reaction) and cDNA (25 μg total RNA reverse transcribed and split into 45–47 reactions) using the KAPA HiFi HotStart ReadyMix PCR kit. In the first gDNA PCR, a primer external to the HDR library was used to prevent amplification of plasmid DNA. cDNA reactions were either primed from exons 16 and 18 (hexamer experiment; Library L) or exons 18 and 20 (Libraries R, R2). After the initial gDNA and cDNA reactions, all PCR products from a single sample were pooled and purified using the QIAquick PCR Purification Kit (Qiagen).

For both cDNA and gDNA reactions, a primer designed to selectively amplify edited molecules bearing the selective PCR site was used either in the first or second reaction. Optimal annealing temperatures for each primer pair were determined via gradient PCR, and negative control reactions were performed using input from HDR library-only transfections to ensure products were derived from edited genomes as opposed to the HDR library. Negative controls failed to amplify for all experiments. Two subsequent PCRs were performed to add sequencing adaptors ('PU1L' and 'PU1R'), sample indices, and flow cell adaptors.

For the *DBR1* experiment, 30 cycles of selective PCR were performed on gDNA (300 ng per reaction) from D5 (3 μg), D8 and D11 (27 μg each). Wells from each sample were pooled, PCR purified, and then re-amplified for 15 additional cycles. The 1,055 bp product was gel-purified (QIAquick Gel Extraction Kit, Qiagen), and two subsequent PCRs were performed to incorporate sequencing and flow cell adaptors before sequencing as above.

After final reactions were purified (AMPure XP beads, Agencourt), paired-end sequencing was performed on all samples with the Illumina MiSeq to quantify gDNA and/or cDNA abundances for each edited haplotype. All primer sequences for reverse transcription, selective PCR, and sequencing library preparation are provided in Supplementary Table 1.

HDR efficiencies were estimated for all experiments via deep sequencing of target loci by performing PCR on 150–300 ng of gDNA using primers external to the region of editing and the selective PCR site. Reported HDR efficiencies were conservatively calculated as the fraction of sequencing reads containing the selective PCR site and bearing at least one variant represented in the HDR library.

Analysis of sequencing data. For quality control, fully overlapping paired-end reads were merged with PEAR[32] (Paired-End reAd mergeR) and discordant pairs were eliminated. By design, the mutagenized region is covered by both the forward and reverse reads on the Illumina platform, resulting in high-confidence calls per site.

For *BRCA1* hexamer reads to be included, the six bases on either side of the hexamer were required to match the reference sequence, and every base call in the hexamer required a quality score of at least Q30. For *BRCA1* whole-exon mutagenesis, the full read was required to be the correct length and match the library consensus sequence outside of the mutagenized region, every base quality score inside the mutagenized region was required to be at least Q30, and no indels were tolerated in alignment with BWA-MEM[33]. cDNA reads not matching any gDNA haplotype with at least 10 reads were eliminated. After normalizing for sequencing coverage, enrichment scores were calculated as cDNA read counts incremented by one pseudocount divided by gDNA reads, calibrated to the wild-type hexamer.

For *DBR1* mutagenesis, reads were subjected to the same requirements of the sequence outside the mutagenized bases matching the consensus and every quality score in the mutagenized region exceeding Q30. Only reads matching programmed haplotypes were analysed, and haplotypes below a D5 relative abundance of 5×10^{-5} of were excluded from analysis. After incrementing all read counts by one pseudocount and dividing by the total number of reads, the abundance of each haplotype on D8 or D11 was divided by the corresponding abundance on D5, and the fold change relative to the wild type sequence was taken to calculate an enrichment score. Based on the bimodal distribution observed in each replicate, mutations with $\log_2$-transformed enrichment scores less than -2 were considered 'deleterious'; otherwise, mutations were considered 'tolerated'. Discordant effects between replicates were defined as mutations 'tolerated' in one replicate but 'deleterious' in the other. Amino acid level enrichment scores were calculated as the median of SNV enrichment scores for programmed edits resulting in the same change (or lack of change, for synonymous edits).

SNV effect size linear modelling and replicate pooling. To determine SNV effects in the *BRCA1* whole-exon experiments, cDNA and gDNA read counts were converted into percentages (number of reads for a given haplotype divided by the total number of reads for a given replicate) after discarding haplotypes with fewer than 10 gDNA reads. Because we had variance in the number of reads for each haplotype, the null expectation of equal variance (σ^2) for each cDNA/gDNA ratio was violated. Because each effect size (y_{ij}) was the average of n_{ij} observations (reads), then $\mathrm{var}(y_{ij}) = \mathrm{var}\ \varepsilon_{ij} = \sigma^2/n_{ij}$, suggesting that the weight for each variable should be n_{ij}. To predict single nucleotide effect sizes across exon 18 of *BRCA1*, we then fit the weighted linear model:

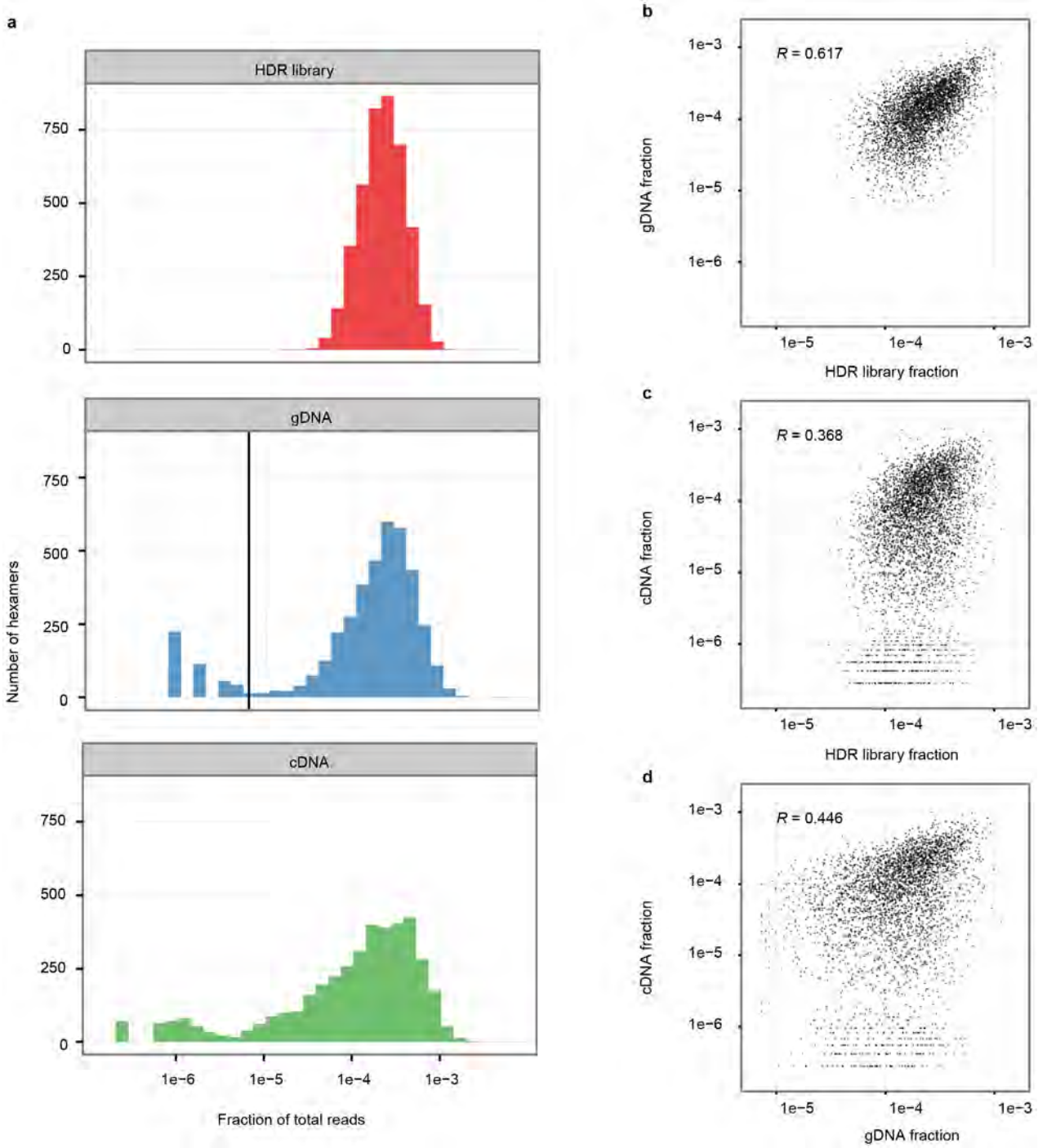
$$y_{ij} = \beta_0 + w_{ij}\beta_{ij}X_{ij}$$

in which y_{ij} is the $\log_2$ enrichment score for a given haplotype, w_{ij} is the number of gDNA reads for a given haplotype, β_{ij} is the effect of nucleotide i at position j relative to the wild-type allele, and X_{ij} is a dummy variable indicating the presence or absence of a particular nucleotide change i at position j relative to the wild type allele. Regression analyses were performed in R 3.0.0 using the lm() function. The resulting coefficients of the model adjusted for the intercepts ($\beta_0 + \beta_{ij}$) were interpreted as effect sizes of the individual SNVs on exon splicing/stability. To merge data across replicates, effect sizes were averaged (including across overlapping bases between libraries L and R in the *BRCA1* exon).

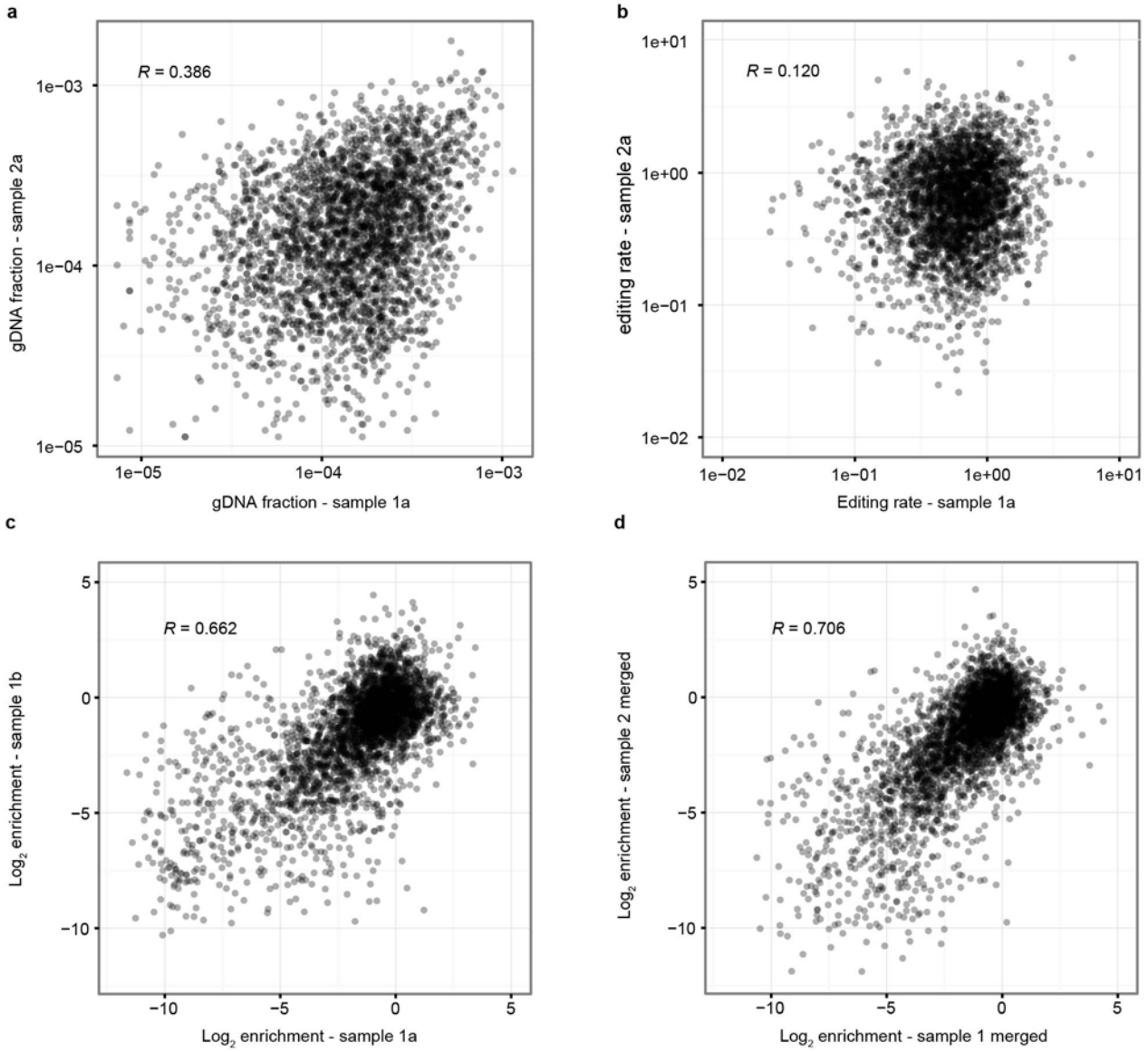
Comparisons to other metrics of functional impact. For comparison to plasmid studies, ESR-seq scores were taken from ref. 14. Hexamers with positive ESR-seq scores are deemed exonic splicing enhancers, whereas negative ESR-seq scores denote exonic splicing silencers. For comparison of *BRCA1* exon 18's SNV effect sizes to an *in silico* method, all SNVs were queried on MutPredSplice's web server (http://mutdb.org/mutpredsplice/submit.htm). MutPredSplice reports a single score estimating the likelihood that a variant will disrupt splicing at any genomic locus. Absolute values of *BRCA1* exon 18 splicing effect sizes were then correlated with MutPredSplice scores to determine concordance between our data and predicted effects on splicing.

For *DBR1*, calculated enrichment scores were compared to BLOSUM62 substitution[34] (obtained from NCBI), PolyPhen-2 (ref. 35), and CADD[22] (PolyPhen-2 and CADD scores obtained from querying genomic coordinates from CADD's precomputed genomic annotations (http://cadd.gs.washington.edu/download).

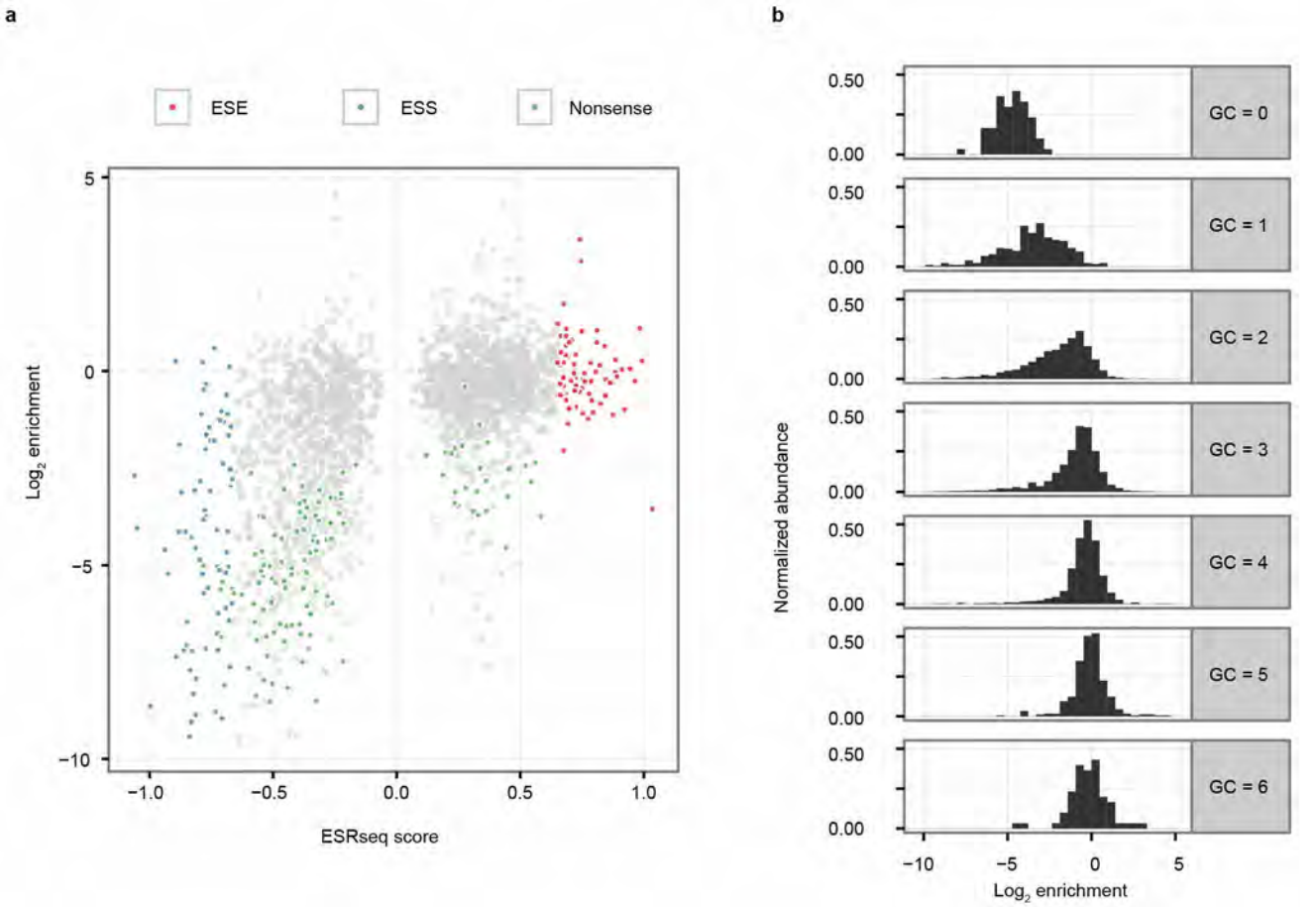
31. Goina, E., Skoko, N. & Pagani, F. Binding of DAZAP1 and hnRNPA1/A2 to an exonic splicing silencer in a natural BRCA1 exon 18 mutant. *Mol. Cell. Biol.* **28,** 3850–3860 (2008).
32. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30,** 614–620 (2014).
33. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
34. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89,** 10915–10919 (1992).
35. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* **7,** 248–249 (2010).



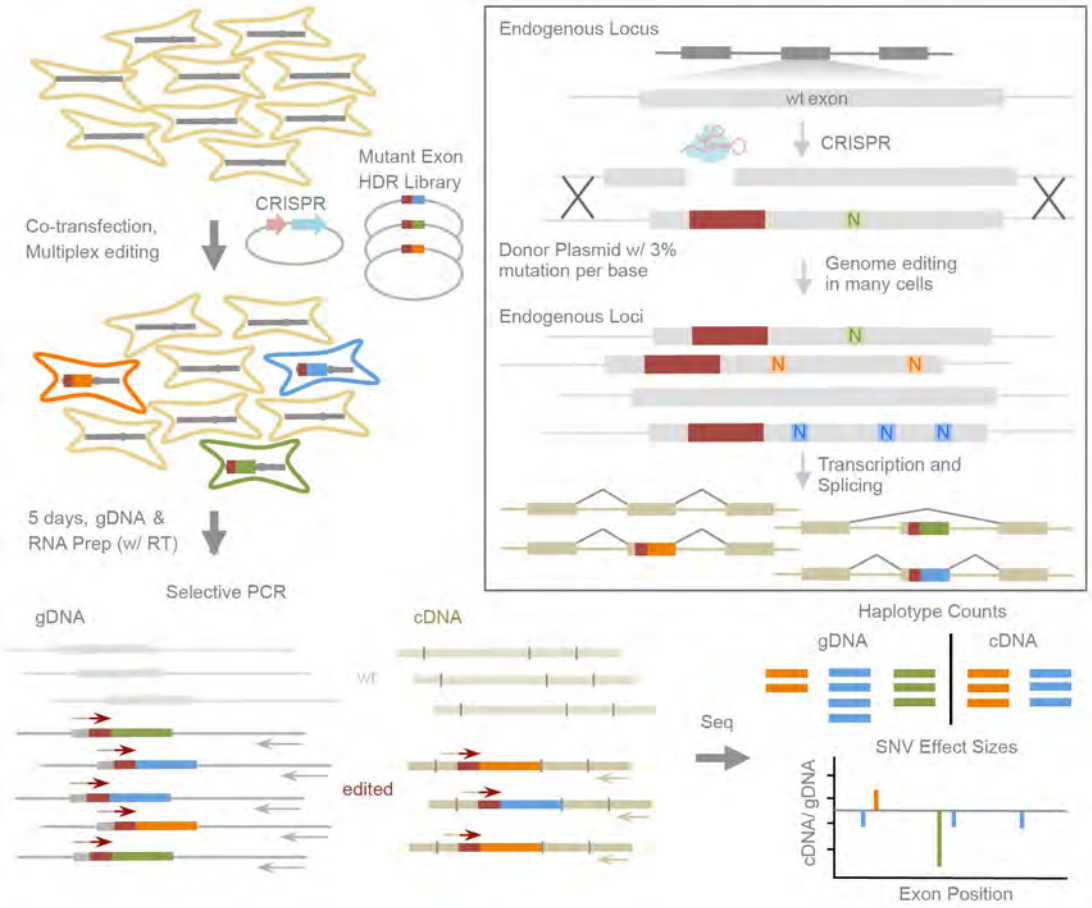
Extended Data Figure 1 | Distributions and pair-wise correlations of hexamer abundances. **a**, The relative abundance of hexamers within the HDR library (red), gDNA (blue), cDNA data (green) are shown for a single experiment. The vertical black line represents our threshold of 10 gDNA reads. **b**–**d**, Scatterplots from a single replicate show pair-wise correlations between sequencing counts for the HDR library, gDNA, and cDNA for hexamers with at least 10 observations in the gDNA library, excluding wild type and control hexamers ($n = 3{,}633$). The HDR library and the gDNA data are most highly correlated (R 95% confidence interval (CI): 0.596–0.636), followed by the gDNA and cDNA (R 95% CI: 0.419–0.471) and the HDR library and cDNA (R 95% CI: 0.341–0.394).



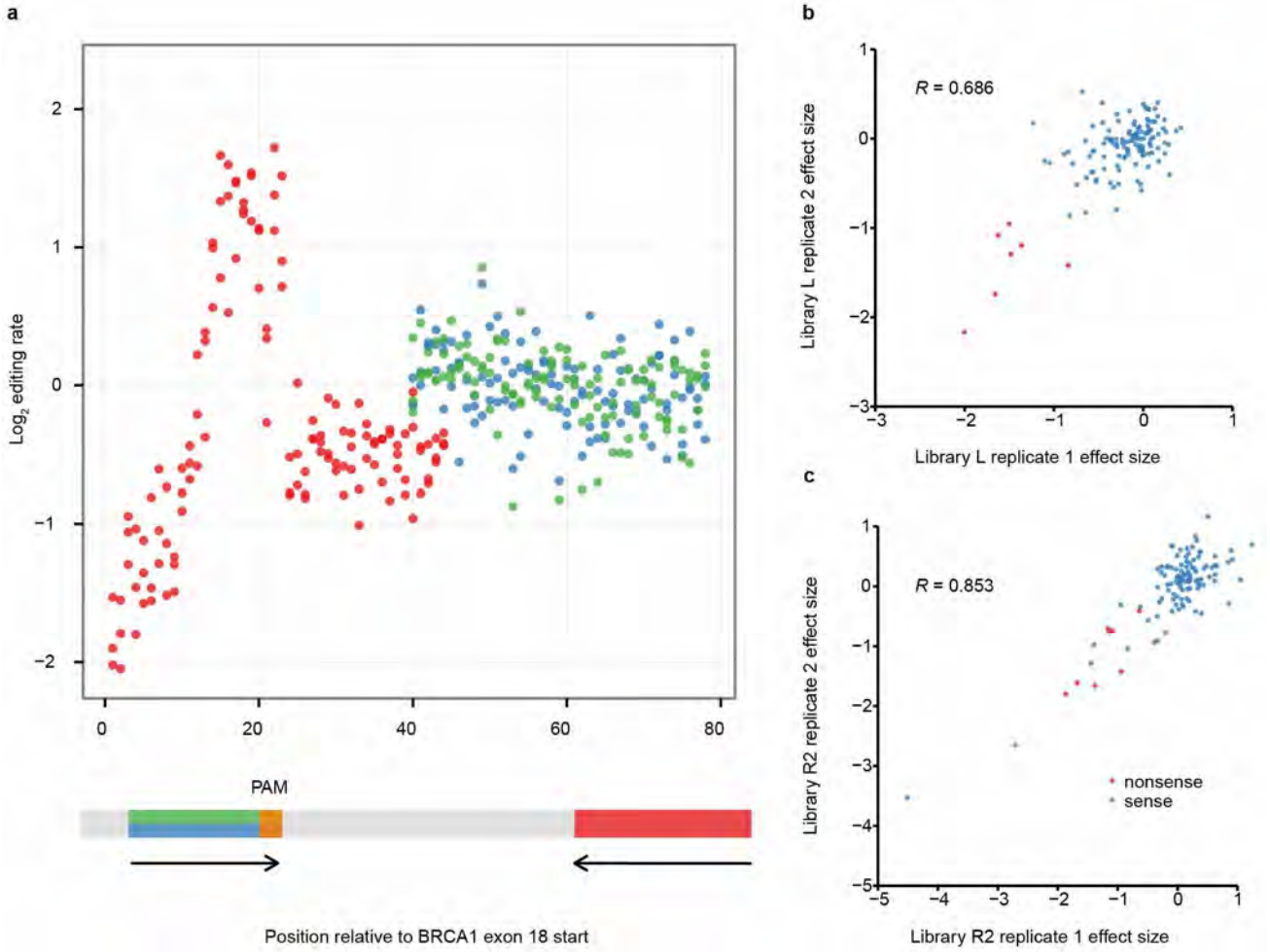
Extended Data Figure 2 | Correlations for hexamer genome editing efficiency and enrichment scores between replicates. **a**, gDNA counts for all hexamers with at least ten reads in each of two gDNA preps from separate transfections with the same HDR library ($n = 2{,}980$) exhibited moderate correlation (R 95% CI: 0.355–0.416). **b**, However, hexamer editing rates, defined as gDNA counts normalized to HDR library counts, were substantially less correlated (R 95% CI: 0.084–0.155), consistent with a hexamer's HDR library abundance contributing more to its gDNA abundance than systematic differences in HDR efficiency secondary to the hexamer sequence itself. **c**, Hexamer enrichment scores for two pools of cells from a single transfection split on D3 were well-correlated (R 95% CI: 0.643–0.681). **d**, Pooling data from cells split on D3 replicates from a single transfection yielded an improved correlation between biological replicates (that is, independent transfections; R 95% CI: 0.690–0.722).



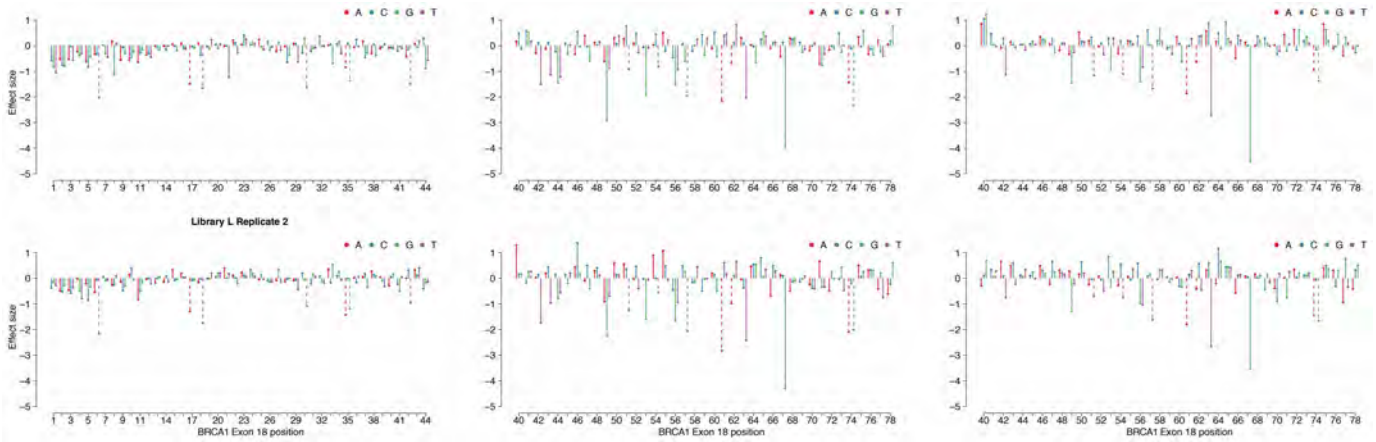
Extended Data Figure 3 | Comparison of genome-based hexamer enrichment scores to plasmid-based hexamer scores. **a**, There was a modest correlation between ESS and ESE hexamers defined by a previous study[14] (*x*-axis) and the enrichment scores calculated here (*y*-axis; Spearman $\rho = 0.524$). The previous study also interrogated hexamers positioned +5 to +10 nucleotides relative to a splice junction, but was plasmid-based rather than genome-based and in the context of different exons. **b**, To reveal effects of GC content on hexamer abundance, histograms display the distribution of enrichment scores for each possible G+C level (0–6). Hexamers containing two or fewer G+C base pairs exhibited broadly lower enrichment scores than hexamers containing three or more G+C base pairs.



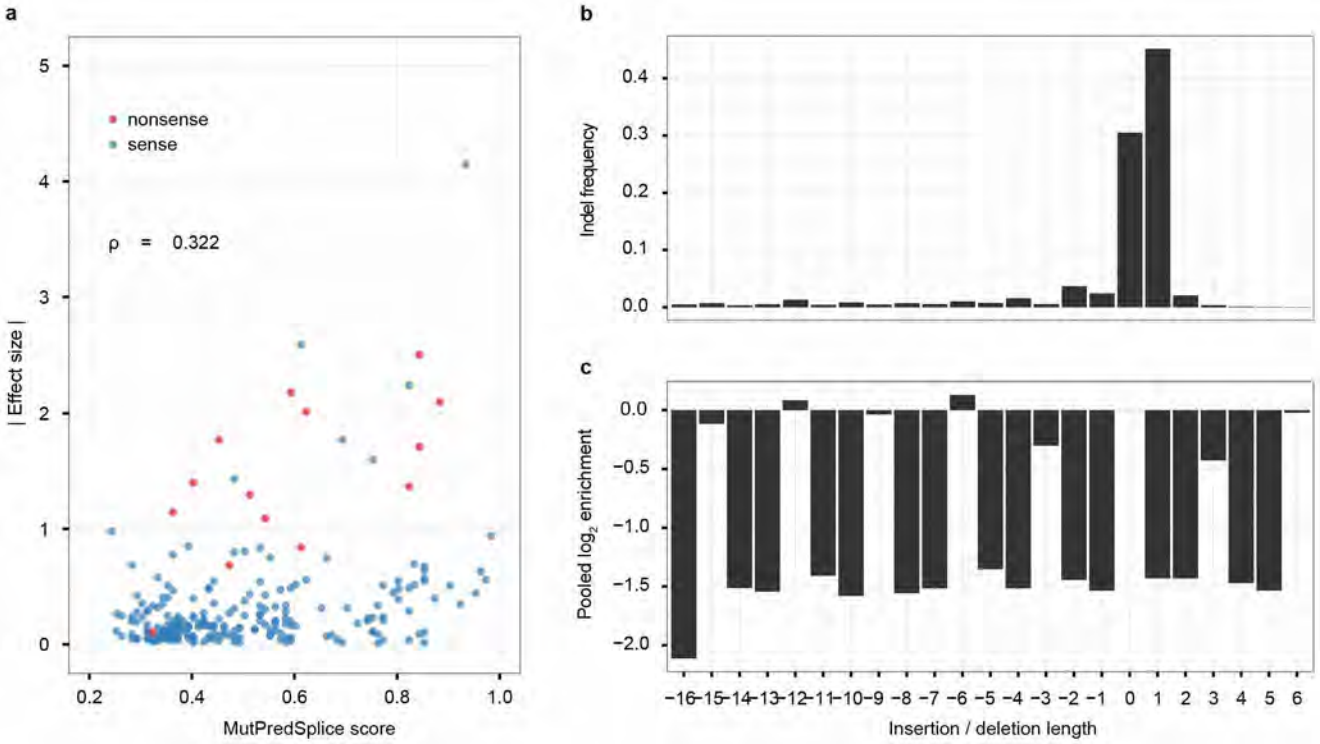
Extended Data Figure 4 | Experimental schematic for genome editing and functional analysis of *BRCA1* exon 18. Cultured cells were co-transfected with a single Cas9-sgRNA construct (CRISPR) and an HDR library. Each HDR library was generated from cloning of an oligonucleotide synthesized with 3% nucleotide degeneracy (97WT:1:1:1) for approximately half of the exon and a selective PCR site introduced to the other (fixed) half of the exon (red). CRISPR-induced HDR integrates mutant exons into the genome. Cells were cultured for five days post-transfection, and then harvested for gDNA and total RNA. After reverse transcription, selective PCR was performed before sequencing the edited pools of gDNA and cDNA. Each exon haplotype's enrichment score was measured by dividing cDNA reads by gDNA reads, and effect sizes for each SNV were calculated via weighted linear regression.



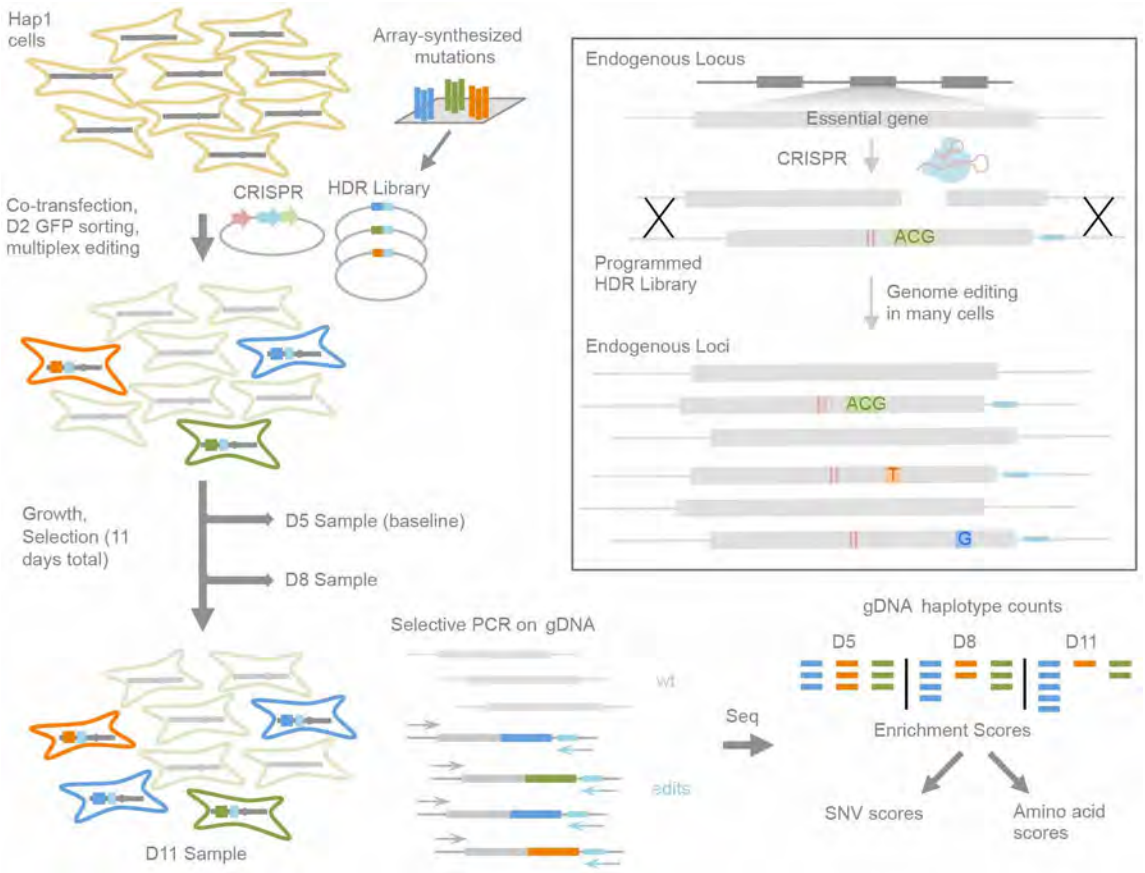
Extended Data Figure 5 | Positional SNV editing rates and replication of effect sizes. a, Editing rates for each SNV in *BRCA1* exon 18 were calculated by dividing each SNV's gDNA sequencing abundance by its HDR library abundance. Editing rates were then plotted across the exon for each library (red = L, blue = R, green = R2) with locations of their selective PCR sites and the CRISPR-targeted PAM illustrated below. For HDR libraries R and R2, there was a subtle decrease in editing rate with increasing distance from the Cas9 cleavage site ($rho^{R} = -0.264$, $p^{R} = 4.1 \times 10^{-3}$; $rho^{R2} = -0.361$, $p^{R2} = 4.8 \times 10^{-5}$). For library L, which allowed re-cutting by not destroying the PAM, there was a sharp peak of editing centred on the Cas9 cleavage site, and a rapid decline in efficiencies in the 5′ direction (further from the 3′ selective PCR handle). **b**–**c**, SNV effect sizes were concordant across biological replicates for libraries R2 (**b**) and L (**c**) (library R shown in Fig. 2). Notably, variants of high effect size scored similarly across independent transfections.



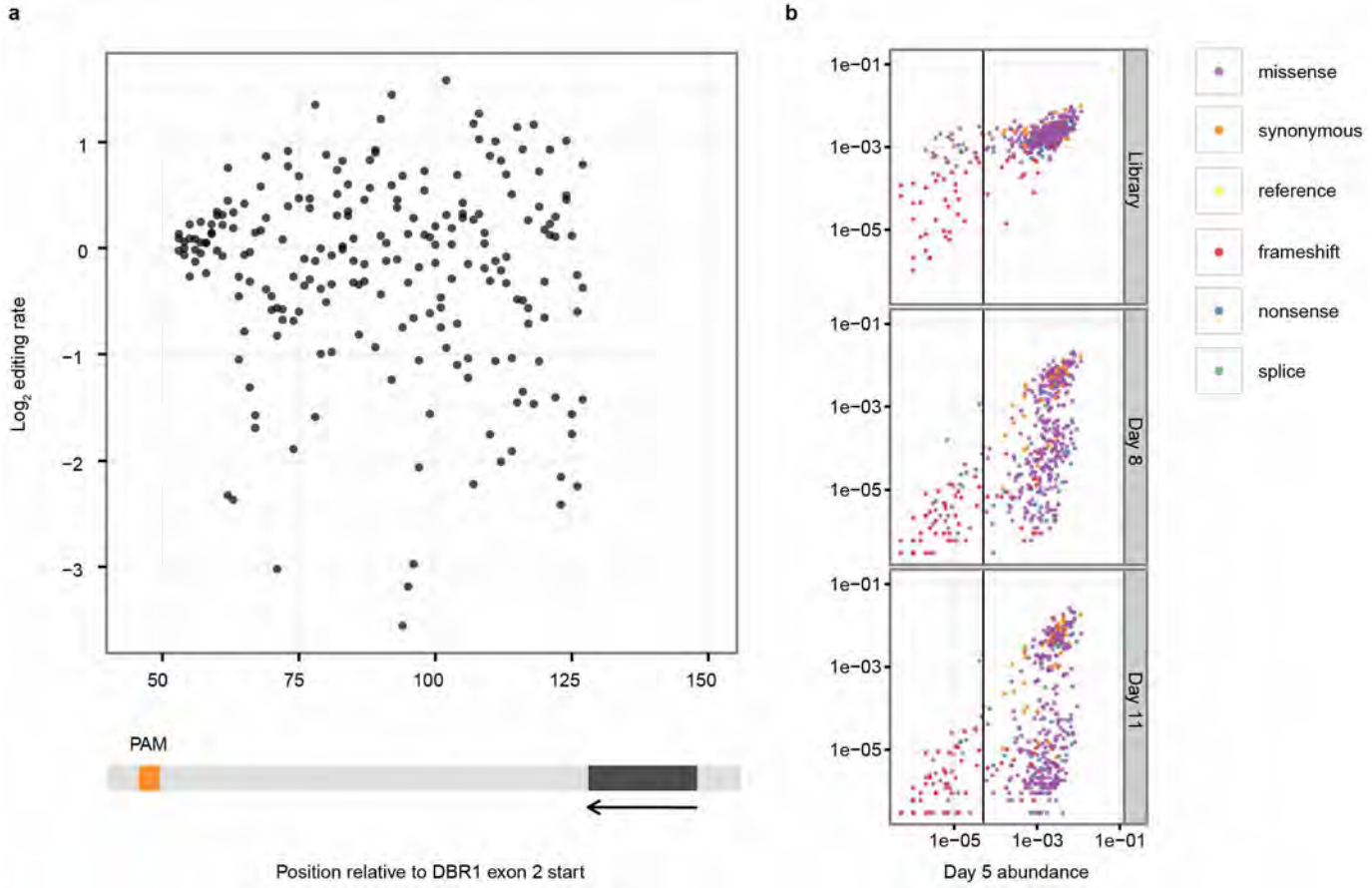
Extended Data Figure 6 | Biological replicate effect size reproducibility for all libraries. Three separate HDR libraries (R, R2, and L) containing 3% nucleotide degeneracy in either half of *BRCA1* exon 18 were introduced to the genome via co-transfection with pCas9-sgBRCA1x18. Enrichment scores were calculated for each haplotype observed at least ten times in the gDNA, and effect sizes of SNVs were determined by weighted linear regression. Effect sizes of individual variants for libraries R2 (left), R (middle), and L (right) were well correlated between biological replicates. Dashed lines represent SNVs that introduce nonsense codons.



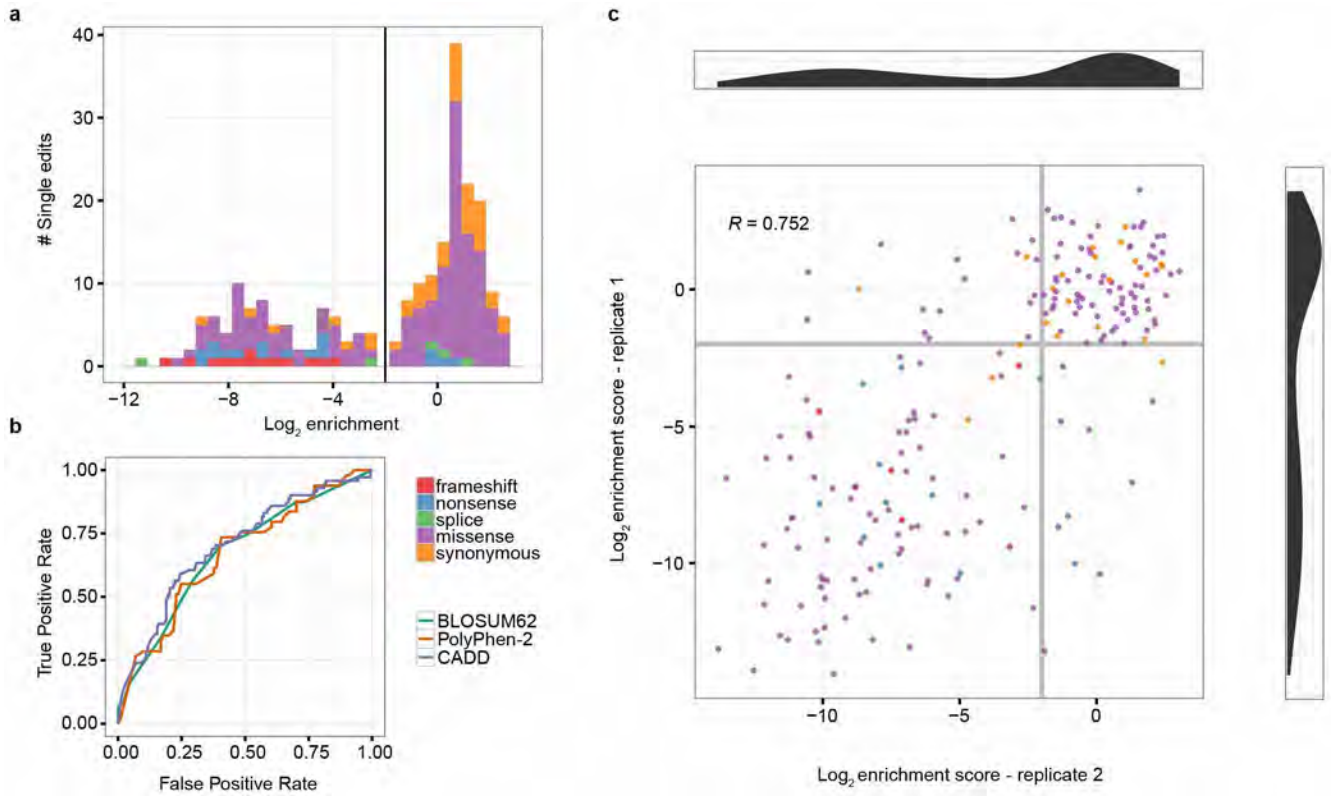
Extended Data Figure 7 | Correlation between effect sizes and predicted disruption of splicing motifs and indel effects. **a**, MutPred Splice[17] was used to predict the functional impact of all 234 single nucleotide substitutions on splicing in *BRCA1* exon 18 (*x*-axis), and these scores were compared to absolute values of our empirically measured effect sizes (*y*-axis; $\rho = 0.322$). Although nonsense variants contributed to this trend, the sense variants with the largest effect sizes generally had high MutPred Splice scores. **b**, For indels observed in gDNA from library 2 (virtually all of which occur at the Cas9 cleavage site), size frequencies are plotted. Indel size = 0 includes all haplotypes with wild type length. **c**, For each indel size, enrichment scores were calculated and normalized to that of the average full length exon. As predicted by nonsense-mediated decay, indels that shift the coding frame were associated with low transcript abundance.



Extended Data Figure 8 | Experimental schematic for saturation genome editing and multiplex functional analysis of *DBR1* exon 2. Hap1 cells were co-transfected with a single Cas9-2A-EGFP-sgRNA construct (CRISPR) and an HDR library cloned from array-synthesized oligonucleotides containing programmed SNVs (orange, blue) and active site codon substitutions (green). The HDR library exon haplotypes also included two synonymous mutations (red) to disrupt PAM and protospacer sequences to prevent Cas9 re-cutting, and a 6 bp selective PCR site (light blue) substituted in the downstream intron. Successfully transfected cells (EGFP+) were selected on D2 by FACS, and cultured. On D5, D8, and D11, samples of cells were taken and selective PCR was performed before targeted sequencing of gDNA. Each haplotype's enrichment score, a measure of the haplotype's fitness in cell culture, was calculated by dividing D8 or D11 abundance by D5 abundance.



Extended Data Figure 9 | *DBR1* editing rates by position and comparison of haplotype abundances between D5 and the HDR library, D8, and D11. **a**, Editing rates for programmed SNVs represented in the *DBR1* gDNA library above threshold ($n = 216$) were calculated by normalizing each SNV's gDNA abundance by its HDR library abundance. Rates are plotted by position, with the locations of the targeted PAM (orange) and selective PCR site (purple) indicated below. The editing rate did not significantly change with position ($P > 0.05$), consistent with positional effects being negated by eliminating re-cutting and performing selective PCR from a distal site. **b**, Scatterplots display the frequencies at which each haplotype was observed in the D5 sample vs the HDR library, D8, and D11 samples. To account for bottlenecking from editing of a limited number of cells in this representative experiment, analysis of individual haplotypes was restricted to those present at frequencies above 5×10^{-5} in the D5 sample ($n = 377$; represented by the vertical line). Selection was evident by the depletion of many haplotypes in D8 and D11 samples.



Extended Data Figure 10 | Performance of computational predictions of deleterious *DBR1* mutations and reproducibility between biological replicates. **a**, D11 enrichment scores from a single experiment were used to empirically define deleterious mutations as those with scores fourfold below wild type (vertical line). **b**, Three *in silico* metrics of functional impairment were tested for their ability to anticipate the deleteriousness of these mutations as indicated by the area under the receiver operating characteristic curve (AUC): BLOSUM62[34] (AUC = 0.672, 214 SNVs), PolyPhen-2[35] (AUC = 0.671, 155 non-synonymous SNVs), and CADD[22] (AUC = 0.701, 214 SNVs). Despite the different approaches of these algorithms, all three exhibited comparably moderate predictive power. **c**, A biological replicate of the *DBR1* experiment was performed and D11 enrichment scores for amino acid substitutions were well correlated (grey lines on scatterplot indicate the 'deleteriousness' threshold of fourfold depletion). The distribution of amino acid level enrichment scores for each experiment is displayed along each axis, reflecting bimodality. Notably, unexpected effects (that is, nonsense mutations scoring as tolerated) were among the relatively small percentage of effects not consistent between replicates.