nature methods

ANALYSIS

https://doi.org/10.1038/s41592-019-0614-5

# Biological plasticity rescues target activity in CRISPR knock outs

Arne H. Smits[1,5], Frederik Ziebell[1,5], Gerard Joberty[2], Nico Zinn[2], William F. Mueller[1], Sandra Clauder-Münster[1], Dirk Eberhard[2], Maria Fälth Savitski[2], Paola Grandi[2], Petra Jakob[1], Anne-Marie Michon[2], Hanice Sun[3], Karen Tessmer[1], Tilmann Bürckstümmer[4], Marcus Bantscheff[2], Lars M. Steinmetz[1,3]*, Gerard Drewes[2]* and Wolfgang Huber[1]*

**Gene knock outs (KOs) are efficiently engineered through CRISPR–Cas9-induced frameshift mutations. While the efficiency of DNA editing is readily verified by DNA sequencing, a systematic understanding of the efficiency of protein elimination has been lacking. Here we devised an experimental strategy combining RNA sequencing and triple-stage mass spectrometry to characterize 193 genetically verified deletions targeting 136 distinct genes generated by CRISPR-induced frameshifts in HAP1 cells. We observed residual protein expression for about one third of the quantified targets, at variable levels from low to original, and identified two causal mechanisms, translation reinitiation leading to N-terminally truncated target proteins or skipping of the edited exon leading to protein isoforms with internal sequence deletions. Detailed analysis of three truncated targets, BRD4, DNMT1 and NGLY1, revealed partial preservation of protein function. Our results imply that systematic characterization of residual protein expression or function in CRISPR–Cas9-generated KO lines is necessary for phenotype interpretation.**

Following the development of CRISPR-based technology[1–3], genome editing has become a widely accessible tool for biological research. The technology is based on an RNA-guided enzyme, such as the nuclease Cas9, bound by a short guide RNA (sgRNA) that provides sequence specificity[4].

One of the most common CRISPR–Cas9 applications is the generation of a genetic KO by the introduction of a frameshift mutation. The Cas9–sgRNA complex is targeted to a specific sequence in the coding region of a gene and cleaves both strands of the DNA[5,6]. The DNA double-stranded break is repaired by non-homologous end joining, an error-prone pathway introducing insertion or deletion mutations that can lead to frameshifts[7] and a premature termination codon (PTC) in the expressed transcript, resulting in nonsense-mediated decay (NMD) of the mRNA and aberrant peptide products that are degraded[8].

Multiple studies have investigated potential off-target effects of CRISPR–Cas9[9–13]. Collectively, these studies report sgRNA-dependent off-target effects at rates that can be comparable to on-target editing. These findings triggered developments to improve CRISPR–Cas9 specificity by the use of Cas9 nickase-variants[12,14,15], improvements in sgRNA design[12,13,16,17] and protein engineering of the nuclease[18,19].

The editing efficiency of CRISPR–Cas9 is commonly assessed on the genomic level. However, many current study designs do not further assess whether the induced frameshift mutation results in the complete loss of protein expression and activity that is expected. Here we systematically characterized the effects of frameshift KO mutations. We studied 193 isogenic KOs in HAP1 cells by quantitative transcriptomics and proteomics. Although all KOs contained frameshifts, about one third of them still expressed the target protein, albeit the majority at reduced levels. Similarly, we detected residual target expression for a CRISPR–Cas9 KO in a K562 cell line. A detailed analysis of three of the KO lines (BRD4, DNMT1 and NGLY1) revealed that the detected protein products were truncated but preserved partial functionality.

## Results

**Systematic characterization of frameshift KO mutations.** We investigated 193 HAP1 lines, in each of which one of 136 genes was targeted by CRISPR–Cas9 to induce a frameshift KO mutation. All lines were derived from the near-haploid human cell line HAP1, which contains a single copy of most genomic loci and is an attractive model system to trace genome-editing events[20–22]. In a first experiment, RNA expression levels of 174 of the mutants and the parental line were measured using 3′ RNA sequencing[23] (3′-RNA-seq; Fig. 1a). In each KO line, we considered the transcripts originating from the targeted gene and calculated the ratio between the expression level in the KO and the parental lines, which we termed the residual level. A residual level of 0 indicates complete elimination, a value of 1 indicates a target transcript level as high as in the parental line. Contrary to our expectation, the 174 residual levels did not concentrate near 0, but covered a wide range. Thus, in many KO lines the mutant mRNA was not strongly reduced. As in all lines the presence of a PTC was confirmed by DNA sequencing, this implies that there was no or weak NMD response to these transcripts. This finding is in line with a recent study that reported high variation in NMD efficiency depending on a complex set of biological factors[24].

Next, we performed an analogous analysis for the target protein levels by tandem mass tag (TMT) isobaric labeling combined with triple-stage mass spectrometry[25] (MS3) (Fig. 1a). We first benchmarked our method with a control experiment in which we spiked *Escherichia coli* into a human protein background in different ratios

[1]European Molecular Biology Laboratory, Heidelberg, Germany. [2]Cellzome, Functional Genomics Research and Development, GlaxoSmithKline, Heidelberg, Germany. [3]Department of Genetics and Stanford Genome Technology Center, Stanford University, Stanford, CA, USA. [4]Aelian Biotechnology, Vienna, Austria. [5]These authors contributed equally: Arne H. Smits, Frederik Ziebell. *e-mail: lars.steinmetz@embl.de; gerard.c.drewes@gsk.com; wolfgang.huber@embl.de

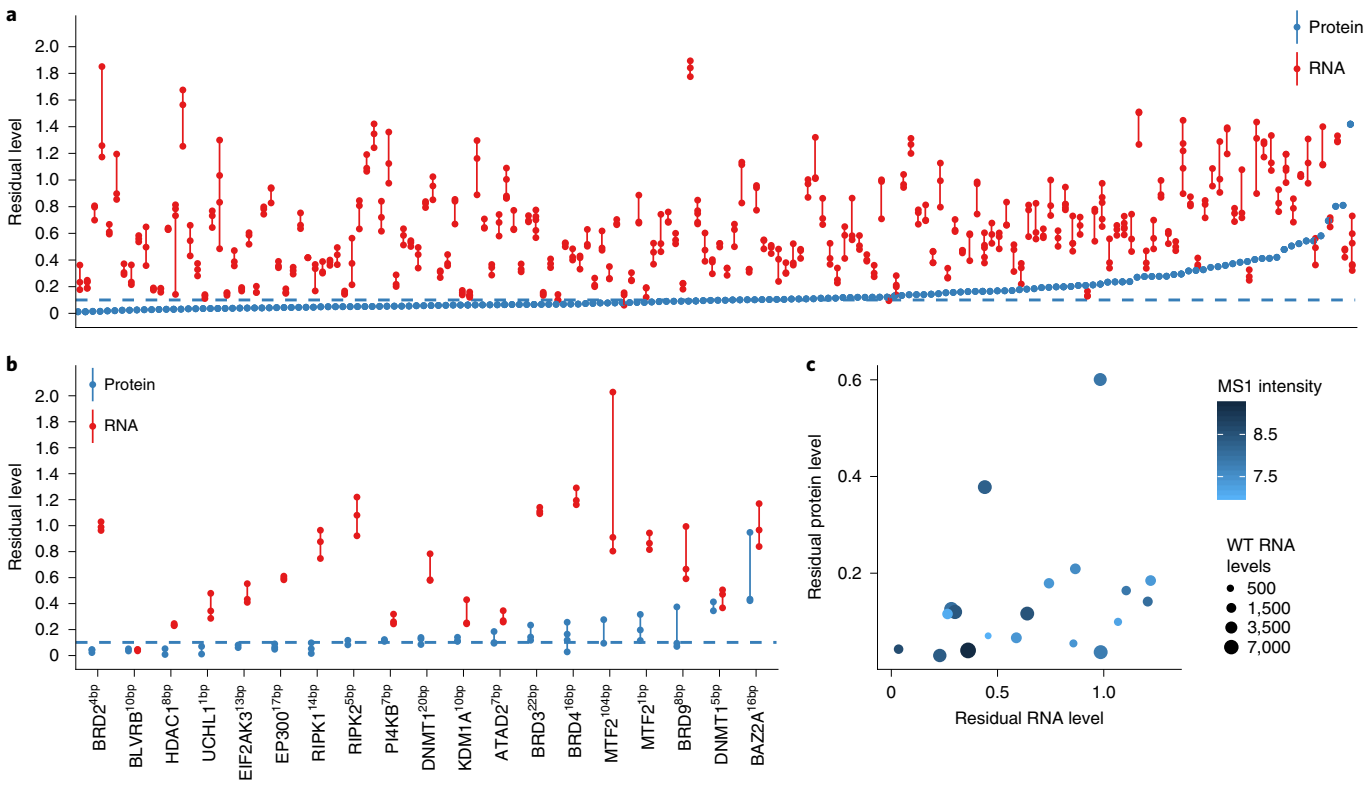


**Fig. 1 | Residual transcript and protein expression in 193 HAP1 cell lines harboring frameshift KO mutations. a**, Residual RNA and protein levels for 174 of the lines. RNA levels were quantified with 3′-RNA-seq with two or more biological replicates (indicated by red points connected by vertical lines). Protein levels were measured with a single MS3 replicate (blue points). **b**, Residual RNA and protein levels of the remaining 19 lines, quantified by whole-transcript RNA-seq and MS3, each with two or more replicates. The data are presented as in **a**. **c**, Scatter plot of RNA and protein expression levels for the 19 KO lines displayed in **b**. The size of dots indicates the wild-type (WT) RNA level and color indicates the protein MS1 signal intensity.

and verified that low protein levels were accurately quantified by the MS3 method (Supplementary Fig. 1a–c). To assess specificity, we considered the distribution of *E. coli* protein quantifications in a human-only sample. While ideally these should all be zero, for 8% (49 of 587), the quantification exceeded 0.1 of the value in the *E. coli*-only sample (Supplementary Fig. 1f and Supplementary Table 1). On this basis, we chose a quantification ratio between KO and parental line of 0.1 as a threshold for calling residual expression and assumed that this would similarly imply a false-positive rate of around 8% (Supplementary Fig. 1d,e).

Surprisingly, about one third of the KO lines displayed residual protein levels above this threshold (Fig. 1a), and there was no correlation between residual RNA and protein levels, indicating that the residual RNA level is not predictive of the residual protein level. We selected an additional set of 19 lines for quantification with whole-transcript RNA-seq and MS3 (Fig. 1b). Each of these quantifications was done with at least two replicates, and replicates were highly consistent (Supplementary Figs. 2 and 3). Similarly to the observation in Fig. 1, there was little correlation between residual RNA and protein levels. Moreover, the residual levels were not dependent on their initial levels before CRISPR modification, as measured by the number of RNA-seq reads and the MS1 intensities (determined by the TOP3 method[26]) of the parental samples (Fig. 1c).

**Residual protein expression owing to exon skipping.** The MTF2$^{104bp}$, PI4KB and BRD3 KO lines showed residual protein levels ranging from 9 to 28% (Fig. 1b). The transcript level of PI4KB was reduced, whereas MTF2$^{104bp}$ and BRD3 displayed levels close to the wild type. Therefore, we assessed whether the mutant lines expressed alternative transcripts that skip the edited exon (Fig. 2a–c).

The *BRD3* transcript in the BRD3 KO line contained exon–exon junction reads spanning from the 5′ untranslated region (UTR) to the second coding exon (11 of 66 reads mapping to this junction) and from the 5′ UTR to the third coding exon (6 of 66 reads) (Fig. 2a). This splicing event removes from the transcript both the start codon and the frameshift mutation. Our data suggest that the 5′ UTR is extended and Met81 or Met125 are used as alternative start codons giving rise to a N-terminally truncated BRD3 protein.

In the MTF2$^{104bp}$ line, the 104-base-pair (bp) insertion and associated frameshift induced a strong alternative splicing event (Fig. 2b). Exon 5, which contains the frameshift mutation, was skipped (60 of 76 reads). Strikingly, this did not occur in another line generated using the same sgRNA, where *MTF2* was mutated with a 1-bp insertion (Fig. 2b). Deletion of exon 5, which is 101 nucleotides long, will lead to a frameshift downstream of the exon. Interestingly, there is an out-of-frame ATG in exon 4 (chromosome 1: 93,114,751–93,114,753, Genome Reference Consortium Human Build 38), which after skipping of exon 5 would result in an in-frame transcript from exon 6 onward. This ATG could explain the observed residual protein expression in the MTF2$^{104\,bp}$ line, by giving rise to a protein isoform that omits the original first 160 amino acids and instead starts with ten alternative amino acids originating from exon 4.

The *PI4KB* frameshift mutation resulted in low-level expression of a splice variant skipping the frameshift-containing exon 5 (Fig. 2c), causing an in-frame deletion of 228 bp (76 amino acids). This splicing event only occurred in ~7% of *PI4KB* transcripts (3 of 42 exon–exon junction reads). This result suggests that the low-level residual protein originates from the alternatively spliced transcript.

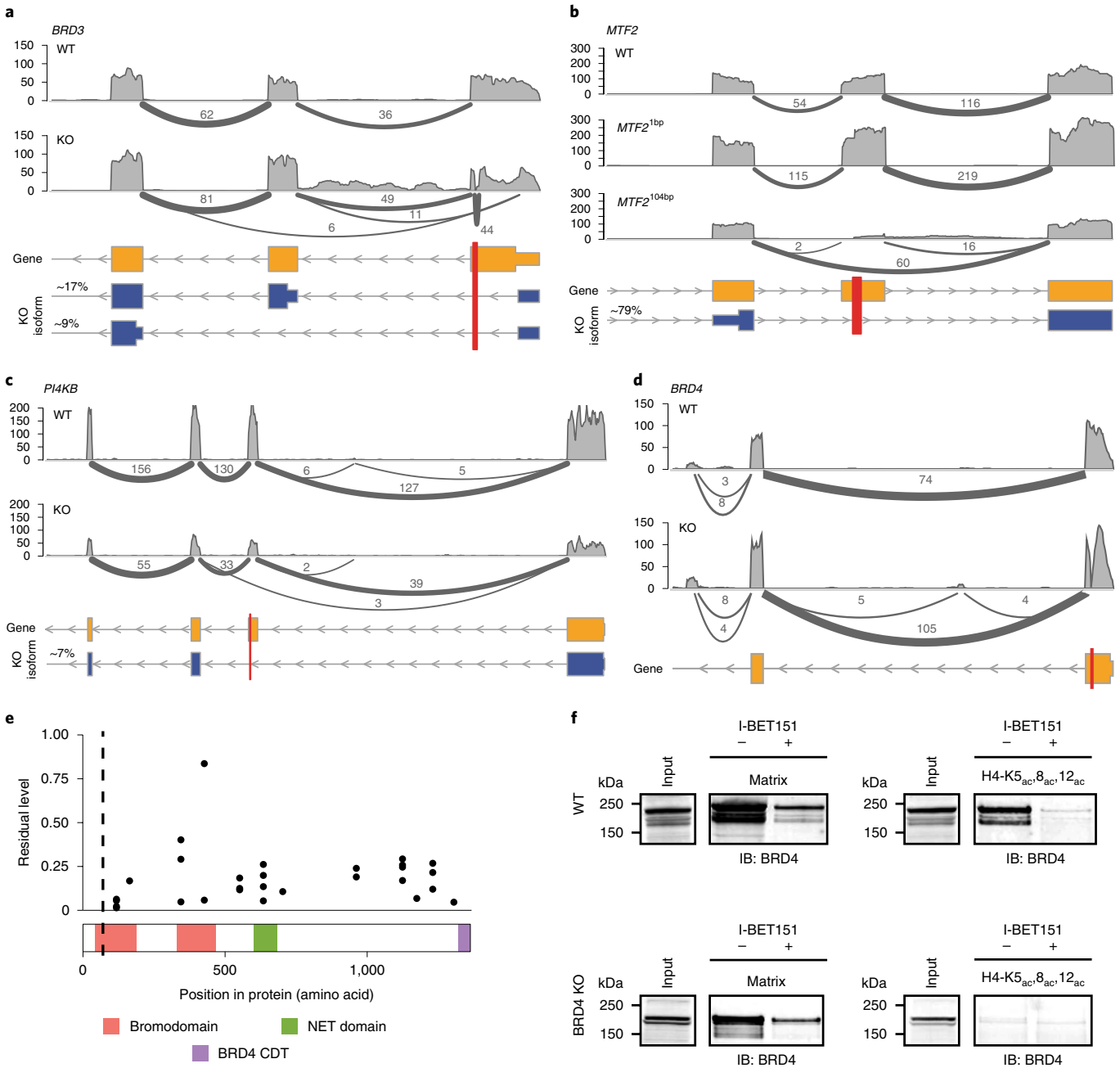


**Fig. 2 | Residual protein expression owing to exon skipping or translation reinitiation. a–d**, Sashimi plots indicate exon usage and splicing of the *BRD3* (**a**), *MTF2* (**b**), *PI4KB* (**c**) and *BRD4* (**d**) transcripts in their respective KO lines. Arches and numbers represent RNA-seq reads at exon–exon junctions. The canonical isoforms are shown in yellow, alternative isoforms identified in the KO lines are shown in blue. Red bars indicate the sgRNA target sites. **e**, Residual peptide levels of BRD4, aggregated per exon. For each biological replicate, the median of all spectra assigned to the amino acid sequence of each exon is depicted. The dashed vertical line depicts the sgRNA target site. The Interpro[40] domain structure of the protein is shown at the bottom. **f**, Binding studies with cell extracts expressing wild-type or truncated BRD4. Affinity pulldowns using beads with immobilized inhibitor (panBET matrix) (left) or biotinylated histone H4ac tails (right) for the parental (top) and BRD4 KO (bottom) line. Matrix, beads with immobilized panBET; I-BET151, panBET bromodomain inhibitor; H4-K5$_{ac}$,8$_{ac}$,12$_{ac}$, acetylated histone H4 tail at positions K5, 8 and 12; IB, immunoblot.

***BRD4* frameshift mutation leads to a partially functional truncated protein by translation reinitiation.** The BRD4 KO line contained a 16-bp deletion in exon 2 (Supplementary Table 2) and showed residual RNA and protein expression (Fig. 1b). However, we did not find evidence for exon skipping (Fig. 2d). To further characterize the residual protein, we looked at residual expression of different regions of BRD4, on the basis of the quantified tryptic peptides. The mutant expressed low levels of the N-terminal region, corresponding to the exons before and around the CRISPR–Cas9 target site (Fig. 2e). The region in the C-terminal direction of the target site displayed residual expression levels higher than 10%. These data suggest N-terminal truncation of BRD4 as a consequence of translation reinitiation[24,27,28]. We confirmed the residual expression of BRD4 by immunoblot, revealing a smaller protein product with an intensity of approximately 15% (Supplementary Fig. 4a). BRD4 immunoprecipitations with two different BRD4 antibodies enriched

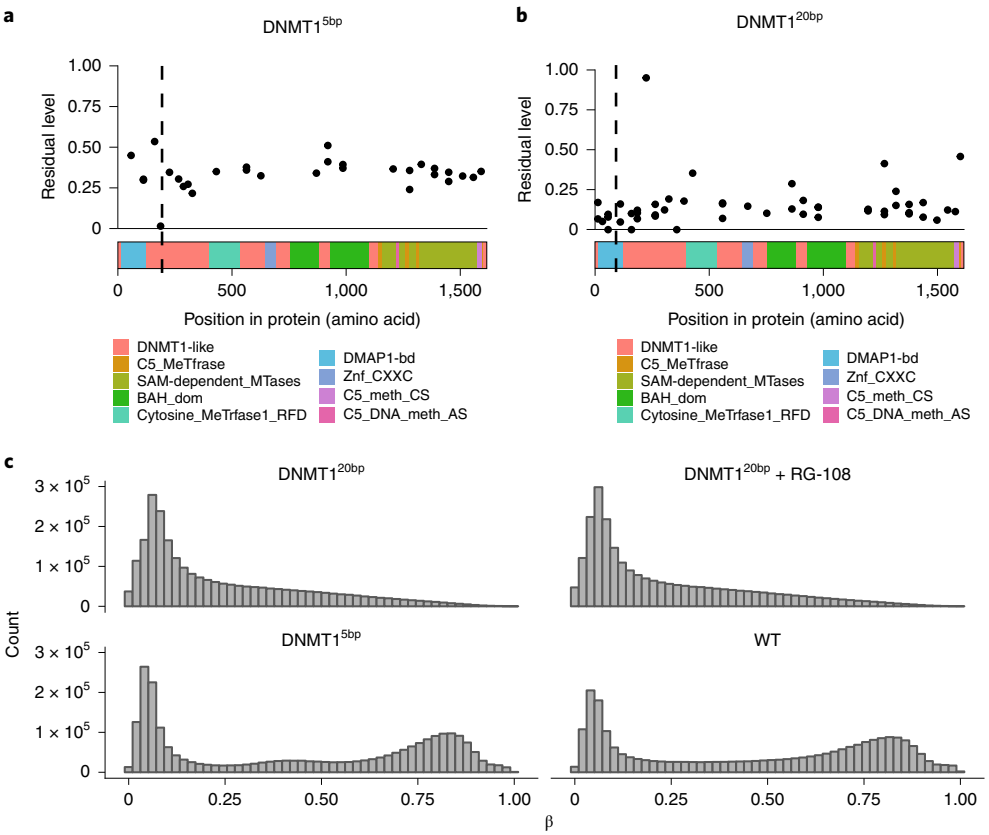


**Fig. 3 | Residual protein expression results in retained activity of DNMT1 in one of the two KO lines. a**,**b**, Residual levels of exons at the protein level of DNMT1$^{5bp}$ (**a**) and DNMT1$^{20bp}$ (**b**). The InterPro[40] domain structure of the protein is shown at the bottom. **c**, Histograms of $\beta$-values from a methylation array experiment. For each pair of probes binding to a given CpG site, $\beta = M/(M+U)$ where $M$ and $U$ are the intensities of the methylated and unmethylated probes, respectively.

both the wild-type and the truncated product, in line with the BRD4 identity of the truncated protein (Supplementary Fig. 4a).

The abundance levels of the quantified peptides and the size of the truncated protein support the explanation that translation of the protein is reinitiated at Met105, Met107 or Met132 (Supplementary Fig. 4b). Notably, Met105 is highly conserved across the bromodomain protein family[29]. For BRD2, a putative isoform is described starting at the methionine corresponding to Met105 in BRD4[30]. Additionally, one of the identified BRD3 truncations (Fig. 2a) starts at Met81, which also corresponds to Met105 in BRD4. Altogether, these properties of Met105 suggest it as a plausible candidate for initiating an N-terminally truncated protein product. The truncated isoform lacks the initial 30 amino acids (41%) of the first bromodomain (BD1), which is crucial for binding of BRD4 to acetylated histone H4 (ref. [29]). To test whether the truncated protein retained any functionality, we performed binding assays with a histone H4 N-terminal peptide and with a small molecule bromodomain inhibitor (I-BET). I-BET binds to the bromodomains of BRD4 and other BET proteins with similar affinity[31,32]. Sepharose beads derivatized with I-BET bound to wild-type and truncated BRD4 as expected, as the truncated protein still comprised the second bromodomain (BD2) (Fig. 2f). By contrast, the triple-acetylated histone H4 tail peptide, which binds to BD1, bound to wild-type but not truncated BRD4 (Fig. 2f). These results confirm the loss of BD1 and the preservation of a functional BD2 in the BRD4 KO.

***DNMT1* frameshift mutation leads to retained DNMT1 activity.** Two DNA methyltransferase 1 (DNMT1) KO lines were assayed by RNA-seq and proteomics (Fig. 1b). The DNMT1$^{5bp}$ line has a 5-bp deletion in exon 6 (Supplementary Table 2) and showed residual peptide levels of about 40% nearly uniformly throughout the gene body (Fig. 3a). Although we observed three reads supporting skipping of the 79-bp-long CRISPR-targeted exon (Supplementary Fig. 5a), this would still lead to a frameshift in the spliced transcript. Instead, the peptide-level expression pattern is better explained by two truncated proteins. The deletion results in a stop codon directly downstream of the CRISPR target site, leading to a C-terminally truncated protein product consisting of the first 188 amino acids. This is validated by the complete absence of the RKPQEESER peptide (amino acids 178–186) in the data for the KO, but not the wild-type line. Although the peptide is still present in the truncated protein product, its new location close to the C-terminal end (RKPQEESERAKstop) results in a modified cleavage context for trypsin. Thus, the peptide is no longer preferentially generated and quantified. A second, N-terminally truncated protein is explained by intron retention. The first quantified peptide downstream of the CRISPR target site is located at the end of exon 8 (peptide EEERDEK, amino acids 234–240). Retention of the intron between exons 7 and 8 creates three in-frame methionines that could be used for translation reinitiation; this hypothesis is in line with annotation of *DNMT1* with twelve transcripts that contain retained introns[33].

The DNMT1$^{20bp}$ line has a 20-bp deletion in exon 4 and here we did not find evidence for excision of the targeted exon (Supplementary Fig. 5b). The nearly uniform residual peptide level of about 10% throughout the gene body (Fig. 3b) is close to the noise threshold established in the MS3 benchmark (Supplementary Fig. 1).

To assess the residual protein activity of both mutant lines, we measured genome-wide methylation levels using methylation

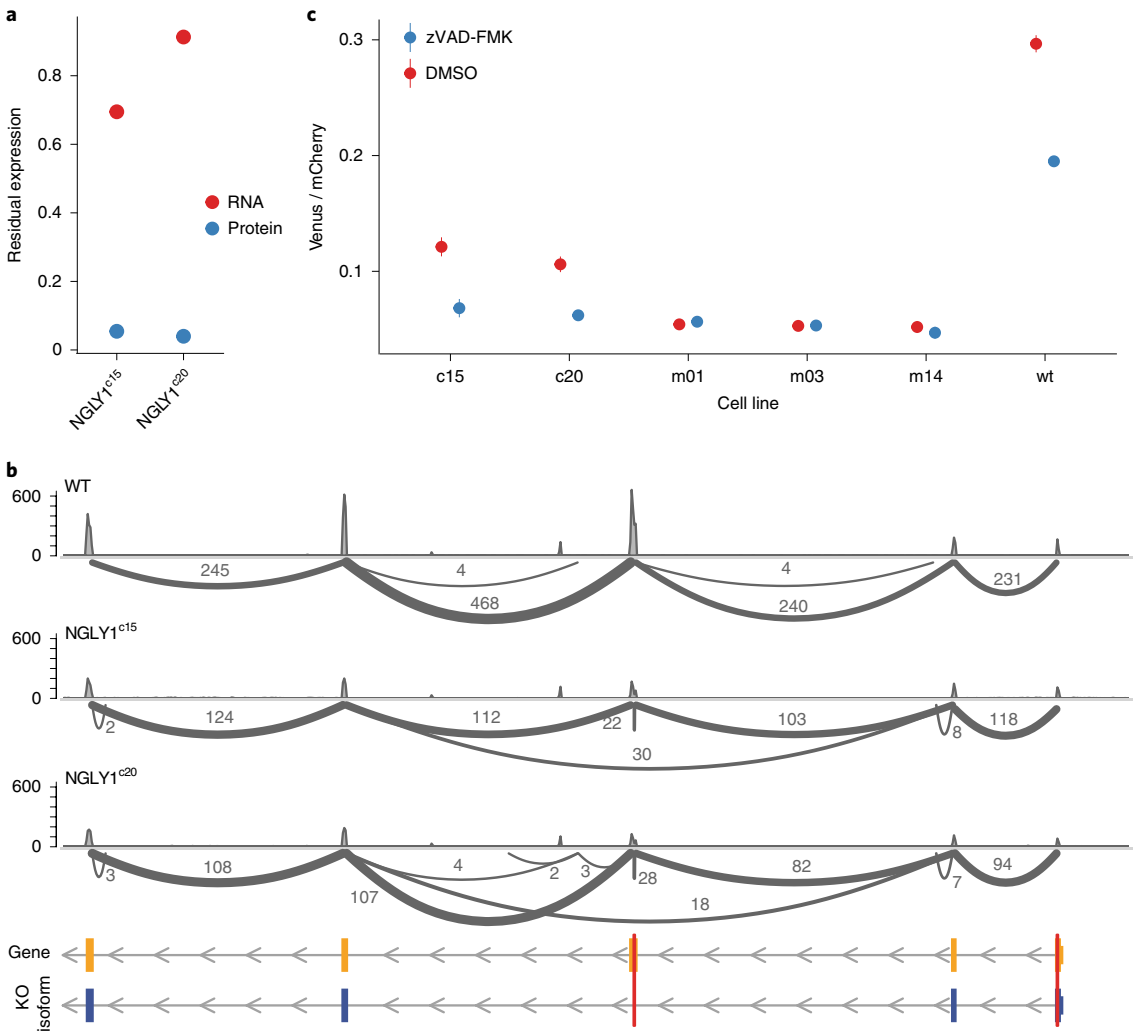


**Fig. 4 | *NGLY1* frameshift mutation in K562 cells results in a partially functional NGLY1 truncation. a**, Residual RNA and protein levels of the two *NGLY1* frameshift mutation K562 lines. Residual RNA levels were determined from a single RNA-seq experiment, residual protein levels were derived from all spectra assigned to the protein in a single mass spectrometry measurement. **b**, Exon usage and splicing of the *NGLY1* transcript in these and the parental lines. **c**, Deglycosylation-dependent Venus fluorescence assay results for the wild-type and *NGLY1* frameshift mutant lines stably expressing the ddVenus reporter. zVAD-FMK, NGLY1 inhibitor; c15 and c20, clones with frameshift mutations in exons 1 and 3; m01, clone with a patient-derived[37] mutation; m03 and m14, clones with a mutation in exon 8.

arrays. DNMT1, the main protein involved in maintenance of DNA methylation, copies pre-existing methylation marks onto hemi-methylated DNA strands during DNA synthesis[34]. The DNMT1$^{5bp}$ line showed methylation levels similar to the wild type (Fig. 3c), consistent with residual protein activity. By contrast, the DNMT1$^{20bp}$ line showed little to no evidence of DNA methylation (Fig. 3c). Treating the DNMT1$^{20bp}$ line with the DNMT1 inhibitor RG-108 did not change its methylation levels, consistent with already complete abrogation of DNMT1 methylation activity. These results show that the DNMT1$^{5bp}$ KO line has residual protein activity.

***NGLY1* frameshift mutation results in a partially functional NGLY1 truncation.** To extend the analysis to a second cell type, we assessed frameshift KO mutations of the *NGLY1* gene, which encodes a de-*N*-glycosylating enzyme, in K562 cells (Fig. 4). Two KO clones termed NGLY1$^{c15}$ and NGLY1$^{c20}$, which each resulted from expansion of a single cell, were isolated. Quantitative mass spectrometry, did not detect residual target protein expression (Fig. 4a). RNA-seq revealed skipping of exon 3, which in both clones contained the frameshift mutation (Fig. 4b). Skipping of this exon results in an in-frame deletion of 246 bp, leading to a truncated protein missing Gly74 to Thr155. Although we did not detect residual protein expression, the expression of the in-frame exon-3-skipping transcript led us to investigate potential residual function of the target enzyme by a deglycosylation assay[35]. Both NGLY1 KO lines showed deglycosylation activity, albeit reduced to 60–65% as compared to wild-type cells (Fig. 4c). Addition of the NGLY1 inhibitor zVAD-FMK[36] further reduced deglycosylation (Fig. 4c). To rule out nonspecific effects of zVAD-FMK, we generated patient-derived[37], termed m01, and frameshift, termed m03 and m14, mutations in *NGLY1* exon 8, that completely abrogated deglycosylation activity (Fig. 4c). Indeed, zVAD-FMK treatment did not further reduce deglycosylation in these clones (Fig. 4c). These results indicate that the cells of the clones NGLY1$^{c15}$ and NGLY1$^{c20}$ had reduced, but measurable, NGLY1-based deglycosylation activity.

## Discussion

We summarize our findings in a model that delineates the possible effects of frameshift KO mutations (Fig. 5). Our first striking observation was the heterogeneity in NMD efficiency in the KO lines.

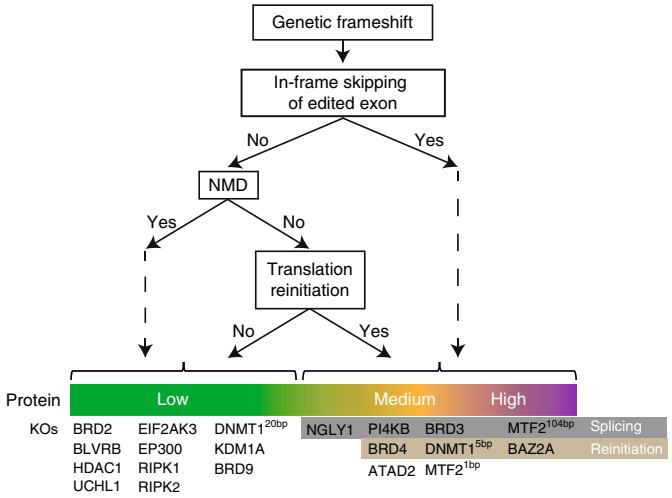


**Fig. 5 | Consequences of CRISPR–Cas9-generated frameshift mutations.** The tree denotes the possible consequences of a frameshift mutation. At the bottom, resulting protein levels are indicated, and we provide a categorization of the 19 HAP1 KO lines shown in Fig. 1b and the NGLY1 KO in K562 cells.

Both alternative splicing, which deletes the PTC from the transcript, and translation reinitiation appeared as causes for incomplete NMD efficiency. However, even in the lines without residual protein expression, NMD efficiency—as assessed by residual RNA level—was variable. This observation is in line with a recent study reporting multiple mechanisms influencing NMD efficiency[24], and indicates that frameshift mutations can lead to full disruption of the protein despite no or weak NMD.

The second striking observation was the residual protein expression in about one third of the KO lines. We identified skipping of the frameshift mutation in the transcript by alternative splicing in four characterized lines. This is in line with previous reports[38,39] and identifies this mechanism as a major source of KO evasion. Moreover, we identified residual protein expression as a consequence of translation reinitiation in three of the KO lines (Fig. 5), in line with a report by Makino et al[28]. The production of truncated proteins by alternative splicing and/or translation reinitiation is an important potential limitation of the CRISPR technology used here. However, it should be possible to at least partially suppress or circumvent these mechanisms by adapting the experimental design, for instance, by using combinations of sgRNAs and/or modifying splicing sites.

We observed that transcript levels of the frame-shifted gene are not predictive of protein levels. The mass-spectrometry-based method applied here has a high signal-to-noise ratio[25] and is able to resolve multiple peptides in different regions of a protein. It is more sensitive and less biased than immunoblotting, which is limited to the epitope of the antibody. As illustrated by the BRD4 KO immunoblot (Supplementary Fig. 4a), proteins that have lower expression levels and truncated proteins are easily overlooked. Therefore, we recommend characterization of KO lines generated using CRISPR–Cas9-induced frameshifts by high-resolution mass spectrometry.

Residual protein expression of the alternatively spliced *NGLY1* frameshift mutation clones was not detected by mass spectrometry, even though residual deglycosylation activity was observed in these lines. This argues for low residual protein expression below the detection threshold of MS3, but sufficient, in the cellular environment, to generate enough enzymatic activity in the time-scale of the applied deglycosylation assay.

In conclusion, our systematic approach enabled us to study the prevalence and nature of adverse effects of CRISPR–Cas9-induced frameshift mutations on target protein levels, including alternative splicing and translation reinitiation. Residual protein function of CRISPR–Cas9 frameshift mutations is an important potential limitation of the CRISPR technology for biological research as well as therapeutic applications.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-019-0614-5.

Received: 3 August 2018; Accepted: 11 September 2019;
Published online: 28 October 2019

## References

1. Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR–Cas9. *Science* **346**, 1258096 (2014).
2. Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR–Cas9 for genome engineering. *Cell* **157**, 1262–1278 (2014).
3. Sander, J. D. & Joung, J. K. CRISPR–Cas systems for genome editing, regulation and targeting. *Nat. Biotechnol.* **32**, 347–355 (2014).
4. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
5. Jinek, M. et al. RNA-programmed genome editing in human cells. *eLife* **2**, e00471 (2013).
6. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
7. Lieber, M. R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem.* **79**, 181–211 (2010).
8. Lykke-Andersen, J. & Bennett, E. J. Protecting the proteome: eukaryotic cotranslational quality control pathways. *J. Cell Biol.* **204**, 467–476 (2014).
9. Fu, Y. et al. High-frequency off-target mutagenesis induced by CRISPR–Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
10. Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
11. Pattanayak, V. et al. High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
12. Cho, S. W. et al. Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.* **24**, 132–141 (2014).
13. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR–Cas nucleases. *Nat. Biotechnol.* **33**, 187–198 (2015).
14. Ran, F. A. et al. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* **154**, 1380–1389 (2013).
15. Shen, B. et al. Efficient genome modification by CRISPR–Cas9 nickase with minimal off-target effects. *Nat. Methods* **11**, 399–402 (2014).
16. Doench, J. G. et al. Rational design of highly active sgRNAs for CRISPR–Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
17. 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).
18. Kleinstiver, B. P. et al. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
19. Slaymaker, I. M. et al. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).
20. Kotecki, M., Reddy, P. S. & Cochran, B. H. Isolation and characterization of a near-haploid human cell line. *Exp. Cell Res.* **252**, 273–280 (1999).
21. Carette, J. E. et al. Haploid genetic screens in human cells identify host factors used by pathogens. *Science* **326**, 1231–1235 (2009).
22. Lackner, D. H et al. A generic strategy for CRISPR–Cas9-mediated gene tagging. *Nat. Commun.* **6**, 10237 (2015).
23. Moll, P., Ante, M., Seitz, A. & Reida, T. QuantSeq 3′ mRNA sequencing for RNA quantification. *Nat. Methods* **11**, 972 (2014).
24. Lindeboom, R. G. H., Supek, F. & Lehner, B. The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nat. Genet.* **48**, 1112–1118 (2016).
25. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8**, 937–940 (2011).
26. Grossmann, J. et al. Implementation and evaluation of relative and absolute quantification in shotgun proteomics with label-free methods. *J. Proteom.* **73**, 1740–1746 (2010).
27. Zhang, J. & Maquat, L. E. Evidence that translation reinitiation abrogates nonsense-mediated mRNA decay in mammalian cells. *EMBO J.* **16**, 826–833 (1997).

28. Makino, S., Fukumura, R. & Gondo, Y. Illegitimate translation causes unexpected gene expression from on-target out-of-frame alleles created by CRISPR–Cas9. *Sci. Rep.* **6**, 39608 (2016).
29. Filippakopoulos, P. et al. Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell* **149**, 214–231 (2012).
30. Bechtel, S et al. The full-ORF clone resource of the German cDNA Consortium. *BMC Genomics* **8**, 399 (2007).
31. Filippakopoulos, P. et al. Selective inhibition of BET bromodomains. *Nature* **468**, 1067–1073 (2010).
32. Dawson, M. A. et al. Inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. *Nature* **478**, 529–533 (2011).
33. Zerbino, D. R. et al. Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2017).
34. Kar, S. et al. An insight into the various regulatory mechanisms modulating human DNA methyltransferase 1 stability and function. *Epigenetics* **7**, 994–1007 (2012).
35. Tomlin, F. M. et al. Inhibition of NGLY1 inactivates the ranscription factor Nrf1 and potentiates proteasome inhibitor cytotoxicity. *ACS Cent. Science* **3**, 1143–1155 (2017).
36. Misaghi, S., Pacold, M. E., Blom, D., Ploegh, H. L. & Korbel, G. A. Using a small molecule inhibitor of peptide: *N*-glycanase to probe its role in glycoprotein turnover. *Chem. Biol.* **11**, 1677–1687 (2004).
37. Enns, G. M. et al. Mutations in NGLY1 cause an inherited disorder of the endoplasmic reticulum-associated degradation pathway. *Genet. Med.* **16**, 751 (2014).
38. Kapahnke, M., Banning, A. & Tikkanen, R. Random splicing of several exons caused by a single base hange in the target exon of CRISPR/Cas9 mediated gene knockout. *Cells* **5**, 45 (2016).
39. Mou, H et al. CRISPR/Cas9-mediated genome editing induces exon skipping by alternative splicing or exon deletion. *Genome Biol.* **18**, 108 (2017).
40. Mitchell, A. L. et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).

## Methods

**Culturing and sequencing of KO clones.** HAP1 KO clones were obtained from Horizon Discovery. Cells were grown in IMDM (Gibco, 21980-032) supplemented with 10% FCS (Gibco, 10270) and collected at 70–90% confluency. Gene target indels from KO cell lines were checked by Sanger sequencing (Sequiserve) from cells with similar passage number as those used to derive transcriptomics and proteomics samples.

**RNA sequencing.** For the quantification of RNA levels, two sequencing techniques were employed. In the technique used for the data shown in Fig. 1a, each KO line was measured by 3′-RNA-seq[23] to a depth of two million reads per sample on average. As this sequencing technique generates reads only at the 3′ end of each transcript molecule, it provides reliable quantification of transcripts at a fraction of the depth that would be needed with shotgun RNA-seq. For the detailed investigation of the subset of 19 KO lines shown in Fig. 1b and the analysis of alternative splicing patterns, KO cell lines were measured with whole-transcript RNA-seq. This technique generates reads throughout all regions of the gene body, and it resulted in a depth of 27 million reads per sample on average.

**3′ RNA sequencing.** Total RNA was isolated from two to three million HAP1 cells using the Direct-zol-96 RNA kit (Zymo Research) according to the manufacturer's protocol. RNA concentration and integrity were determined on a Fragment Analyzer (Advanced Analytical). cDNA libraries for 3′ mRNA sequencing were prepared from 500 ng of total RNA using the QuantSeq 3′ mRNA-Seq Library Prep Kit FWD for Illumina (Lexogen) according to the manufacturer's protocol. For library amplification, 12 PCR cycles were applied and 45–48 libraries were pooled for sequencing on HiSeq 2000 or HiSeq 2500 instruments (Illumina) using 50-bp single-end reads.

**Whole-transcript RNA sequencing.** Total RNA was isolated from two to three million HAP1 cells using the Direct-zol-96 RNA kit (Zymo Research) according to the manufacturer's protocol. RNA concentration and integrity were determined on a Fragment Analyzer (Advanced Analytical). Preparation of cDNA libraries was done with the TruSeq RNA Library Prep Kit v2 from Illumina according to the manufacturer's instructions. cDNA was amplified by 15 PCR cycles and 10–11 samples were pooled before sequencing of 100-bp single-end reads on a HiSeq 4000 instrument (Illumina).

**Data analysis.** RNA-seq reads were aligned with STAR v.2.5.3a[41] to the Human Reference Genome GRCh37. Mapping of reads to the genome and the presence of frameshift mutations were investigated with IGV[42]. Differential expression analysis was performed using the R/Bioconductor package DESeq2[43]. Sashimi plots were generated using the R/Bioconductor package Gviz[44] and gene models were obtained from ENSEMBL using biomaRt[45].

**MS3 proteomics.** Cells were collected by scraping into cold PBS, washed and snap-frozen in liquid nitrogen. Lysis was performed in 2% SDS for 3 min at 95 °C in a thermomixer (Thermo Fisher Scientific), followed by digestion of DNA with benzonase at 37 °C for 1.5 h. Lysate was cleared by centrifugation and the protein concentration in the supernatant was determined by BCA assay. Proteins were reduced by Tris(2-carboxyethyl)phosphine (TCEP) and alkylated with chloroacetamide, separated on 4–12% NuPAGE (Invitrogen) and stained with colloidal Coomassie before proceeding to trypsin digestion. Alternatively, proteins were digested following a modified solid-phase-enhanced sample preparation protocol[46]. In brief, proteins in 2% SDS were bound to Paramagnetic beads (SeraMag Speed beads, GE Healthcare, 45152105050250 and 651521050502) on filter plates (Multiscreen, Merck-Millipore, 10675743) and filled to 50% ethanol. After four washes of the beads with 200 µl of 70% ethanol, beads were resuspended in trypsin and LysC in 0.1 mM HEPES, pH 8.5 containing TCEP and chloroacetamide and incubated at room temperature overnight. Peptides were collected and after lyophilization subjected to TMT labeling.

*E. coli* spiked into human protein background were produced by a tryptic digest of an *E. coli* as described above. Peptides were pooled and separated at defined ratios before TMT labeling. The human background consisted of pooled tryptic digests from human cell lysates labeled with TMT.

Peptides were labeled with isobaric mass tags (TMT10; Thermo Fisher Scientific) using the 10-plex TMT reagents, enabling relative quantification of ten conditions in a single experiment[47,48]. The labeling reaction was performed in 40 mM triethylammonium bicarbonate or 100 mM HEPES, pH 8.5 at 22 °C and quenched with glycine. Labeled peptide extracts were combined to a single sample per experiment and lyophilized.

**Sample preparation for mass spectrometry, prefractionation.** Lyophilized samples were resuspended in 1.25% ammonia in water. The whole sample was injected onto a precolumn (2.1 mm × 10 mm, C18, 3.5 µm (Xbridge, Waters)) at a flow rate of 15 µl $min^{-1}$. Separation was done at 40 µl $min^{-1}$ on a reversed-phase column (1 mm × 150 mm, C18, 3.5 µm (Xbridge, Waters)) with a 115-min-long gradient ranging from 97% buffer A (1.25% ammonia in water) to 60% B (1.25% ammonia in 70% acetonitrile in water)[49].

**Liquid chromatography with tandem mass spectrometry.** Samples were dried in vacuo and resuspended in 0.05% trifluoroacetic acid in water. Half of the sample was injected into an Ultimate3000 nanoRLSC (Dionex) coupled to a Q-Exactive or Orbitrap Fusion Lumos (Thermo Fisher Scientific). Peptides were separated on custom-made 50 cm × 100 µm (ID) reversed-phase columns (Reprosil) at 55 °C. Gradient elution was performed from 2% acetonitrile to 30% acetonitrile in 0.1% formic acid and 3.5% DMSO over 1 or 2 h. Q-Exactive plus mass spectrometers were operated with a data-dependent top ten method. Mass spectra were acquired using a resolution of 70.000 and an ion target of $3 \times 10^6$. Higher energy collisional dissociation (HCD) scans were performed with 33% normalized collision energy (NCE) at 35.000 resolution (at an $m/z$ of 200), and the ion target setting was set to $2 \times 10^5$ to avoid ion coalescence[48]. The instruments were operated with Tune 2.2 or 2.3 and Xcalibur 2.7 or 3.0.63.

Synchronous precursor selection MS3 experiments were performed on an Orbitrap Fusion Lumos operated with fixed cycle time of 2 s. Mass spectra were acquired at a resolution of 60.000 and an ion target of $4 \times 10^5$. Tandem mass spectra were acquired in the linear ion trap following HCD fragmentation performed at 38% NCE and an ion target of $1 \times 10^4$. MS3 were performed synchronously on the five most abundant MS2 signals in the $m/z$ range of 400–2,000 using HCD at 70% NCE in the Orbitrap at a resolution of 30.000 and an ion target of $1 \times 10^5$. The instrument was operated with Tune 2.1.1565.23 and Xcalibur 4.0.27.10. Targeted data acquisition in combination with synchronous precursor selection was performed according to Savitski et al.[50], employing the settings described above.

**Peptide and protein identification.** Mascot 2.4 (Matrix Science) was used for protein identification by using a ten parts per million mass tolerance for peptide precursors and 20 mD (HCD) mass tolerance for fragment ions. The search database consisted of a UniProt protein sequence database combined with a decoy version of this database created by using scripts supplied by Matrix Science. Carbamidomethylation of cysteine residues and TMT modification of lysine residues were set as fixed modifications. Methionine oxidation, and N-terminal acetylation of proteins and TMT modification of peptide N-termini were set as variable modifications. We accepted protein identifications as follows. (1) For single spectrum to sequence assignments, we required this assignment to be the best match and a minimum Mascot score of 31 and a 10× difference of this assignment over the next best assignment. On the basis of these criteria, the decoy search results indicated <1% false-discovery rate (FDR). (2) For multiple spectrum to sequence assignments, and using the same parameters, the decoy search results indicated <0.1% FDR. Quantified proteins were required to contain at least two unique peptide matches. FDR for quantified proteins was <0.1%.

**Isobaric-mass-tag-based quantification.** Reporter ion intensities were read from raw data and multiplied with ion accumulation times (the unit is milliseconds) so as to yield a measure proportional to the number of ions; this measure is referred to as ion area[51]. Spectra matching to peptides were filtered according to the following criteria: mascot ion score > 15, signal-to-background of the precursor ion > 4 and signal-to-interference > 0.5 for MS2-based quantification[51]. Fold changes were corrected for isotope purity as described and adjusted for interference caused by co-elution of nearby isobaric peaks as estimated by the signal-to-interference measure[51]. MS3-based quantification was performed identically but excluding the adjustment for interference caused by co-elution of nearby isobaric peaks. Protein quantification was derived from individual spectra matching to distinct peptides by using a sum-based bootstrap algorithm; 95% confidence intervals were calculated for all protein fold changes that were quantified with more than three spectra[51] for MS2-based quantification. An automatic outlier removal procedure was performed on proteins with at least five quantified peptide-spectrum matches. In brief, each peptide-spectrum match, and the respective quantification values, were sequentially excluded from the fold change calculation; in cases where the confidence interval was then improved by at least 30%, the given quantification values were permanently excluded from further analyses.

**Data analysis.** Raw mass spectra were analyzed by Mascot and IsobarQuant[52]. Downstream analyses were performed with R and Bioconductor[53]. TMT reporter intensities were used to calculate the residual KO expression, that is, the intensity ratio of the KO line over wild-type HAP1, per spectrum. Missing reporter intensities were imputed with zeros. Residual protein levels of exons were calculated by taking the median of all spectra uniquely mapping to every exon of the protein of interest. Exon annotation was obtained from ENSEMBL using biomaRt[45] and translated to the corresponding amino acid numbers. The position of the spectra was determined as the middle of the identified peptide. Protein domain annotation was obtained from ENSEMBL using biomaRt[45]. All plots were generated using ggplot2 (ref. [54]).

**BRD4 immunoprecipitation.** BRD4 immunoprecipitation was performed using two antibodies. Rabbit anti-BRD4 amino acids 1312–1362 (Bethyl Laboratories, A301985A) and rabbit anti-BRD4 (Abcam ab128874). One hundred microliters of AminoLink resin (Thermo Fisher Scientific) was functionalized with antibody. HAP1 WT or HAP1 BRD4 KO lysate (100 µl) was added and incubated for 2 h at 4 °C. Beads were subsequently washed and eluted by adding SDS sample buffer.

BRD4 products were detected by immunoblotting using rabbit anti-BRD4 (Santa Cruz, sc-48772) antibody.

**Inhibitor and histone tail pulldowns.** Inhibitor and histone tail affinity enrichment experiments were performed as described previously[32]. Five microliters of NeutrAvidin beads (Thermo Scientific) were functionalized with either 0.02 mM panBET inhibitor (GSK923121) or 0.01 mM H4(1–21)-K5$_{ac}$,K8$_{ac}$,K12$_{ac}$. HAP1 WT or HAP1 BRD4 KO lysate (100 µl) was added and incubated for 2 h at 4 °C. Competition was performed by adding a final concentration of 10 µM I-BET151 or DMSO as control. Beads were subsequently washed and proteins were eluted by adding SDS sample buffer. Immunoblotting was performed using rabbit anti-BRD4 (Santa Cruz, sc-48772) antibody.

**Methylation analysis.** A wild-type HAP1, DNMT1$^{5bp}$ and DNMT1$^{20bp}$ KO clone were seeded at $0.2 \times 10^6$ cells per well in triplicates (six wells for the DNMT1$^{20bp}$ clone). The next day, cells were treated with 0.1% DMSO except for three wells of the DNMT1$^{20bp}$ clone that were treated with the DNMT1 inhibitor RG-108 (GW546419X) at 5 µM (0.1% DMSO final) for 24 h. The following day, cells were collected and genomic DNA was isolated using the Blood and Cell Culture DNA Mini kit from Qiagen according to the manufacturer's instructions. Samples were analyzed using the Infinium MethylationEPIC BeadChip (Illumina) according to the manufacturer's instructions. Raw signal intensities obtained from IDAT files were processed with the R/Bioconductor[53] package minfi[55]. Probes containing single-nucleotide polymorphisms were removed and subsequently $\beta$-values were calculated from raw intensities.

**NGLY1 KO generation.** K562 cells were transfected with two sgRNAs targeting exon 1 and 3 of *NGLY1* and Cas9 plasmid (lentiCas9-Blast) by nucleofection according to the manufacturer's protocol (Nucleofector, Lonza). LentiGuide-Puro (Addgene plasmid, 52963) and lentiCas9-Blast (Addgene plasmid, 52962) were gifts from F. Zhang (Broad Institute of MIT and Harvard, USA)[56]. Positive transfected cells were selected and enriched with puromycin (4 µg µl$^{-1}$) and blasticitidin (3.5 µg µl$^{-1}$) for 15 d. NGLY1 knockout was confirmed in the cell population by immunoblotting using the anti-NGLY1 antibody from Sigma (HPA036825). Single GFP$^-$ cells were sorted into a 96-well plate and expanded. NGLY1 KO clones were confirmed by Sanger sequencing.

The following sgRNAs were used: exon 1 gRNA, CTTGGAGGCCTCCAAAA; exon 3 gRNA, TCTGCTACTTCTCTCTA; and exon 8 gRNA, GCAAACATGAAGAGGTGATT.

**Deglycosylation assay.** The deglycosylation-dependent Venus fluorescence assay was performed as described previously[35]. Parental and NGLY1 KO clones were infected or transfected with ddVenus reporter and mCherry expression constructs. Cells were treated with DMSO or 20 µM zVAD-FMK for 6 h at 37 °C. Subsequently, mean fluorescence of Venus and mCherry were measured by flow cytometry on a BD LSR Fortessa using the HTS attachment. mCherry levels were used to control for transfection efficiency and cell size.

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

## Data availability

The transcriptomics and proteomics data of the 19 KO lines shown in Fig. 1b and the NGLY1 KO lines were deposited to publicly available repositories. The mass spectrometry proteomics data were deposited to the ProteomeXchange Consortium via the PRIDE[57] partner repository with the dataset identifier PXD010335. RNA-seq data were deposited in the ArrayExpress database at EMBL-EBI (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-7061. The data for the remaining 174 KO lines shown in Fig. 1a are available from the corresponding author upon request.

## References


41. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
42. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24 (2011).
43. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
44. Hahne, F. & Ivanek, R. in *Statistical Genomics: Methods and Protocols* (Eds. Mathé, E & Davis, S) 335–351 (Humana Press, 2016).
45. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
46. Moggridge, S., Sorensen, P. H., Morin, G. B. & Hughes, C. S. Extending the compatibility of the SP3 paramagnetic bead processing approach for proteomics. *J. Proteome Res.* **17**, 1730–1740 (2018).
47. Werner, T. et al. High-resolution enabled TMT 8-plexing. *Anal. Chem.* **84**, 7188–7194 (2012).
48. Werner, T. et al. Ion coalescence of neutron encoded TMT 10-plex reporter ions. *Anal. Chem.* **86**, 3594–3601 (2014).
49. Savitski, M. M. et al. Multiplexed proteome dynamics profiling reveals mechanisms controlling protein homeostasis. *Cell* **173**, 260–274 (2018).
50. Savitski, M. M. et al. Targeted data acquisition for improved reproducibility and robustness of proteomic mass spectrometry assays. *J. Am. Soc. Mass Spectrom.* **21**, 1668–1679 (2010).
51. Savitski, M. M. et al. Measuring and managing ratio compression for accurate iTRAQ/TMT quantification. *J. Proteome Res.* **12**, 3586–3598 (2013).
52. Franken, H. et al. Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. *Nat. Protoc.* **10**, 1567–1593 (2015).
53. Huber, W. et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
54. Wickham, H. *Elegant Graphics for Data Analysis* (Springer-Verlag, New York, 2009).
55. Fortin, J.-P., Triche, T. J. Jr & Hansen, K. D. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* **33**, 558–560 (2016).
56. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
57. Vizcaino, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–D456 (2015).


## Acknowledgements

We would like to thank D. Pavlinic, F. Jung and V. Benes for RNA-seq; J. Stuhlfauth and team for cell banking; M. Boesche and team for mass spectrometry analyses; S. Shimamura for cell culture and sample preparation; and the microarray unit of the DKFZ Genomics and Proteomics Core Facility for providing the Illumina Human Methylation arrays and related services. A.H.S. was supported by a fellowship from the EMBL Interdisciplinary Postdoc (EIPOD) Programme under a grant from the Marie Sklodowska-Curie Actions COFUND (664726). The NGLY1 work was supported by the Grace Science Foundation.

## Author contributions

A.H.S. and F.Z. analyzed the data. A.H.S. and D.E. performed transcriptomics experiments. N.Z. performed mass spectrometry measurements. G.J. resequenced cell lines and analyzed BRD4 and DNMT1 truncation data. W.F.M., K.T. and H.S. performed all NGLY1 KO experiments and initial analyses. P.J. and S.C.-J. created and validated the NGLY1 KO cell lines and established the functional reporter in K562 to test NGLY1 functionality. A.-M.M. performed the BRD4 functional experiments. P.G., T.B., M.F.S., M.B., L.M.S., W.H. and G.D. supervised the work. A.H.S., F.Z., W.H. and G.D. wrote the manuscript with input from all authors.

## Competing interests

G.J., N.Z., D.E., M.F.S., P.G., M.B., G.D. are employees and/or shareholders of Cellzome and GlaxoSmithKline. T.B. was an employee of Horizon Genomics GmbH.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41592-019-0614-5.

**Correspondence and requests for materials** should be addressed to L.M.S., G.D. or W.H.

**Peer review information** Nicole Rusk was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

**Reprints and permissions information** is available at www.nature.com/reprints.

nature research

Corresponding author(s): Lars Steinmetz, Gerard Drewes, Wolfgang Huber

Last updated by author(s): Aug 26, 2019

# 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.

## Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | A statement on 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<br>*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 statistical parameters 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<br>*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 |

*Our web collection on statistics for biologists contains articles on many of the points above.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | RNA-seq: STAR v2.5.3a for alignment, proteomics: IsobarQuant v1.0.0 and Mascot v2.4 for analysis of raw mass spectra. |
| Data analysis | RNA-seq data analysis: DESeq2 v1.20 (R/Bioconductor) for library-size normalization and vst-transformation, Gviz v1.28 (R/Bioconductor) for visualization of aligned reads and IGV v2.5.0 (Broad Institute) for their visual inspection. biomaRt (R/Bioconductor) for gene annotation.<br>Methylation data analysis: minfi v1.29.4 (R/Bioconductor) for calculation of beta values |

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. 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

The transcriptomics and proteomics data shown in Figure 1b have been deposited in ArrayExpress under accession number E-MTAB-7061 and the ProteomeXchange Consortium with identifier XXXXXXX. The datasets shown in Figure 1a are available from the corresponding author on reasonable request.

# Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-flat.pdf

# Life sciences study design

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

| | |
|---|---|
| Sample size | No sample size was determined, since the aim of the study was to screen a large collection of knockouts and report residual RNA and protein levels for all of them. We used a single replicate for explorative analysis (Fig 1a) and two to three replicates for validation analysis (Fig 1b and c). This represents the minimum to estimate accuracy of the methods used and at the same time allowed for the screening of a large collection of KO lines. |
| Data exclusions | One HDAC1 KO line was removed from the analysis. The line was measured with two biological replicates (proteomics) and the two replicates showed very strong differences of residual protein levels (0.6 vs 0.01), that have not been seen in any other KO line with replicates. Since the study focused on reporting reproducible patterns that can be seen upon CRISPR-induced frameshift deletions, we decided to remove the KO line from the analysis. Exclusion criteria were not pre-established. |
| Replication | For the subset of KO lines shown in Figure 1b, the correlation of replicates is overall very high and is shown in Supplementary Figures S2 and S3. All attempts of replication were successful. |
| Randomization | For transcriptomics and proteomics 10 (deep-seq, proteomics) or 48 (shallow-seq) samples were multiplexed. Conditions were partially randomized between different multiplexing runs and in addition, each run contained at least one internal control to allow comparison across runs. Partial randomization was performed on the labels (oligos for RNA-sequencing and TMT for proteomics) attached to the samples and dependent on the availability of the cell cultures. |
| Blinding | Blinding was not relevant to our study, since the aim was to screen a large collection of knockouts and evaluate residual RNA and protein levels. |

# Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☐ | ☒ Antibodies |
| ☐ | ☒ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☒ | ☐ Animals and other organisms |
| ☒ | ☐ Human research participants |
| ☒ | ☐ Clinical data |

## Methods

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

## Antibodies

| | |
|---|---|
| Antibodies used | Rabbit anti-BRD4: Abcam, catalog #ab128874, lot #GR275920-1, conc. 1 µg antibody/µg bead<br>Rabbit anti-BRD4 amino acids 1312-1362: Bethyl Laboratories, catalog #A301-985A100, lot #5, conc. 1 µg antibody/µg bead<br>Rabbit anti-BRD4: Santa Cruz #sc-48772, catalog #sc-48772, lot #C2014, conc. 1/100<br>anti-NGLY1 antibody: Sigma, catalog #HPA036825, conc. 1/1000 |
| Validation | Abcam anti-BRD4: According to the manufacturer's website, the antibody has been "Validated in WB, IP, IHC, Flow Cyt, ICC/IF and tested in Mouse, Rat, Human. Cited in 36 publication(s)."<br>Bethyl Laboratories anti-BRD4: According to the manufacturer's website, BRD4 has been IPed with the antibody and appeared as a single band, not present in the IgG control.<br>Santa Cruz anti-BRD4: According to the manufacturer, BRD4 has been detected as single band in whole cell extract.<br>anti-NGLY1: The antibody has been validated by the Human Protein Atlas: https://www.proteinatlas.org/ENSG00000151092-NGLY1/antibody |

## Eukaryotic cell lines

Policy information about cell lines

| | |
|---|---|
| Cell line source(s) | HAP1 WT and KO cell lines were obtained from Horizon Discovery. The K562 cell line was obtained from SIGMA, product ID 89121407. |
| Authentication | Cells are obtained from and curated by Horizon Discovery. According to the manufacturer's website "all HAP1 knockout cell lines are validated by PCR amplification and Sanger Sequencing to confirm the mutation at the genomic level." In addition, we checked a subset of clones by Sanger sequencing for the presence of a frameshift mutation. These clones were pure and had the frameshift described by Horizon Discovery. |
| Mycoplasma contamination | Cell lines were tested by Horizon Discovery for contamination. According to the manufacturer's website: "We follow strict bio-banking procedures and our cells are regularly tested for contamination, including mycoplasma." |
| Commonly misidentified lines (See ICLAC register) | No commonly misidentified lines according to ICLAC. |