Highly parallel genome variant engineering with CRISPR-Cas9

Meru J. Sadhu1,2,3,4,8*, Joshua S. Bloom1,2,3,4,8*, Laura Day1,2,3, Jake J. Siegel1,2,3, Sriram Kosuri4,5,6,7 and Leonid Kruglyak1,2,3,4*

Understanding the functional effects of DNA sequence variants is of critical importance for studies of basic biology, evolution, and medical genetics; however, measuring these effects in a high-throughput manner is a major challenge. One promising avenue is precise editing with the CRISPR-Cas9 system, which allows for generation of DNA double-strand breaks (DSBs) at genomic sites matching the targeting sequence of a guide RNA (gRNA). Recent studies have used CRISPR libraries to generate many frameshift mutations genome wide through faulty repair of CRISPR-directed breaks by nonhomologous end joining (NHEJ). Here, we developed a CRISPR-library-based approach for highly efficient and precise genome-wide variant engineering. We used our method to examine the functional consequences of premature-termination codons (PTCs) at different locations within all annotated essential genes in yeast. We found that most PTCs were highly deleterious unless they occurred close to the 3′ end of the gene and did not affect an annotated protein domain. Unexpectedly, we discovered that some putatively essential genes are dispensable, whereas others have large dispensable regions. This approach can be used to profile the effects of large classes of variants in a high-throughput manner.

Precise gene editing by CRISPR-Cas9 requires using a DNA template for homology-directed repair (HDR) and incorporating the desired sequence variants encoded in the template into the genomic locus. Generating many uniquely edited cells in parallel thus requires each cell to receive the correct gRNA-repair template pair. We devised an approach that accomplishes such pairing by encoding gRNA targeting sequences and their corresponding repair templates in cis on oligonucleotides generated in bulk through high-throughput synthesis. These oligonucleotide libraries are then used to generate pools of plasmids pairing the two components for delivery into yeast cells (Supplementary Fig. 1). We used this approach to understand the consequences of one important class of genetic variants, PTCs.

PTCs interrupt the open reading frames (ORFs) of protein-coding genes. Such mutations are generally expected to have strong deleterious effects, either by abrogating or by changing the functions of the encoded proteins or by causing mRNA degradation through the nonsense-mediated decay (NMD) surveillance pathway. More than 10% of annotated pathogenic human variants are PTCs. Nonetheless, understanding of the detrimental effects of PTCs is incomplete, particularly when they occur near the 3′ ends of genes. Such mutations may not shorten the encoded proteins sufficiently to affect their function, and they often escape NMD.

We first tested gene editing using a plasmid-encoded paired gRNA and repair template (Fig. 1a) by targeting eight specific PTCs to the Saccharomyces cerevisiae genome. S. cerevisiae has a naturally high propensity to repair DSBs through HDR, which we enhanced by using a haploid yeast strain in which NHEJ is abolished by a deletion of the NEJ1 gene (Δnej1; Supplementary Table 1). For each targeted mutation, we sequenced the corresponding genomic locus in thousands of transformed yeast cells. In all eight cases, the desired mutation was present in >95% of sequencing reads, thus demonstrating the high efficiency of this strategy (Table 1). We also tested editing in wild-type diploid yeast, in which NHEJ is inactive, and observed high efficiency at most sites (Supplementary Table 2). None of the sites showed a high rate of insertion or deletion (indel) formation in either the Δnej1 or diploid strains, in agreement with NHEJ being inactive.

We next scaled up the approach by using large-scale oligonucleotide synthesis to generate a pool of more than 10,000 distinct paired gRNA-repair template plasmids (Supplementary Fig. 1). These plasmids targeted PTCs to different sites in 1,034 yeast genes considered essential for viability. Each gene was targeted at ten sites, chosen with a preference for sites closer to the 3′ end (Supplementary Fig. 2). Targeted PTCs were represented by multiple independent barcoded plasmids. We transformed haploid Δnej1 yeast in bulk with plasmid pools in two independent replicate experiments. After inducing Cas9 expression, we collected millions of surviving transformed cells every 24 h for 4 d (Fig. 1b). PTCs that disrupt the function of genes essential for viability were expected to drop out of the pool over time, whereas those that did not were expected to persist.

We determined the abundance of each barcoded edit-directing plasmid at each time point through bulk sequencing, then computed a ‘PTC tolerance score’ based on the persistence of the barcoded plasmids over the duration of the time-course experiment (Methods). PTC tolerance scores from the replicate experiments were correlated at r = 0.6 (P < 2 x 10−16) (Supplementary Fig. 3). As controls, we used a set of 90 ‘dubious ORFs’, which were originally annotated as genes but later reclassified because of a lack of conservation and ascribable function. As expected, PTCs in essential genes were much less tolerated than those in dubious ORFs (Wilcoxon rank-sum test, P < 2 x 10−16) (Fig. 1c). As a further control, 71 sites
in essential genes were targeted with two plasmids that had the same gRNA but different repair templates, only one of which introduced a PTC. Plasmids that introduced a PTC were significantly less tolerated (Supplementary Fig. 3) (paired t-test $t = 6.5, P = 8 \times 10^{-9}$), thus showing that the observed phenotypic effects were predominantly due to specific introduction of the desired mutations rather than repair-template-independent Cas9 activities.

One possibility that may explain the observed PTC intolerance is that most truncations of essential genes may fatally disrupt the function of the encoded proteins. Another possibility is that NMD may remove most transcripts carrying PTCs, and this removal is fatal in the case of PTCs in essential genes. We tested these alternatives by introducing PTCs in an NMD-deficient strain. PTCs in this strain were similarly deleterious (Supplementary Fig. 4) ($\chi^2 = 1.66, P = 0.20$) (Supplementary Table 3), thus suggesting that protein truncation, rather than degradation of transcripts via NMD, explains the observed PTC intolerance.

Although most PTCs in annotated essential genes were highly deleterious, some appeared to be tolerated. We examined the relationship between tolerance scores and locations of PTCs. PTCs were generally deleterious when they were located more than 27 codons away from the gene end (Fig. 1d). Within the 27 terminal codons, the tolerance scores rose toward the 3’ end. PTCs were also more tolerated if they did not interrupt or remove an annotated protein domain ($\chi^2 = 317.2, P = 5.86 \times 10^{-71}$) (Supplementary Fig. 5 and Supplementary Table 3). PTCs that disrupted protein domains tended to be deleterious even when they were located close to gene ends. Evolutionary conservation of the truncated region among related yeast species also had an effect on PTC tolerance ($\chi^2 = 49.8, P = 1.66 \times 10^{-12}$) (Supplementary Fig. 5).

---

**Fig. 1 | Measuring the effects of engineered PTCs in essential genes.** 
**a,** Schematic of pairing of CRISPR gRNA and repair template on plasmids. 
**b,** Experimental design. After Cas9 induction, DNA was extracted every 24 h. At each time point, edit-directing plasmids were quantified by sequencing. 
**c,** Tolerance scores for $n = 8,353$ PTCs targeting essential genes and $n = 694$ PTCs targeting dubious ORFs are shown, with overlaid box plots. The center line of each box corresponds to the median value; the top and bottom of the box span the first quartile to the third quartile; and the whiskers reach to either the most extreme values or 1.5 x the interquartile range. $P < 2 \times 10^{-16}$, two-sided Wilcoxon rank-sum test. 
**d,** Scatter plot of PTC tolerance scores versus distance in codons from the 3’ ends of essential genes. The thick blue line shows a segmented regression fit. Vertical blue lines indicate the 95% confidence interval for the boundary between the segments. The segmented regression was fit on PTC tolerance scores for $n = 7,583$ PTCs that were within 500 codons of the 3’ end of a gene.
Fig. 2 | PTC tolerance of genes. a, Gene tolerance scores for essential genes and dubious ORFs, shown as a violin plot displaying the individual data points. b, Analysis of conditionally essential genes in yeast tetrads. Each vertical set of four colonies corresponds to the four haploid meiotic products from a diploid yeast strain. Each diploid was heterozygous for a deletion mutation of interest and for an interacting mutation. Haploid colonies carrying the deletion of interest are highlighted in red or blue according to their genotype at the interacting locus. Absence of a visible colony (first five panels) indicates a lethal interaction; small colonies (last panel) indicate an interaction causing poor growth. n = 10 tetrads were examined for the ssy1Δ, ptr3Δ, ssy5Δ, and fur1Δ interactions; n = 6 tetrads were examined for the shr3Δ interaction.

We built a model to more precisely delineate dispensable 3’ ends of essential genes. Although our experiment was not designed to comprehensively rule out the existence of small dispensable C termini, interestingly, 517 genes did not appear to tolerate any tested PTCs, even those located very close to the ends (Supplementary Fig. 6 and Supplementary Note 1). In contrast to these highly PTC-intolerant genes, 101 genes tolerated five or more PTCs, thus suggesting that these genes have large dispensable C termini (Supplementary Fig. 6). We computed the overall tolerance of PTCs for each gene and observed considerable variation among genes (Fig. 2a). A Gene Ontology enrichment analysis15 showed that genes encoding proteins with catalytic activity were significantly less PTC tolerant than other genes (Kolmogorov–Smirnov test, Bonferroni-corrected $P = 0.0024$) (Supplementary Table 4 and Supplementary Fig. 7), whereas genes with functions relating to mRNA splicing and processing were significantly more PTC tolerant (Kolmogorov–Smirnov test, Bonferroni-corrected $P = 0.0017$).

To better understand why some genes annotated as essential could tolerate many PTCs, we closely examined the 16 most PTC-tolerant genes (Fig. 2a). We found that three of those genes had been misannotated as essential because their deletion disrupts the function of a nearby essential gene, thus illustrating the value of PTC introduction for characterization of gene essentiality (Supplementary Fig. 8, Fig. 2a, and Supplementary Note 1). PTC-tolerant genes also included Ssy1, Ptr3, and Ssy5, which encode the three members of the SPS (Ssy1–Ptr3–Ssy5) plasma membrane amino acid–sensor system14, as well as SHR3, which is required for SPS cell-surface localization15. Defects in SPS function compromise leucine uptake, and the strain originally used to determine which genes are essential is deficient in leucine biosynthesis and consequently requires leucine uptake, thus explaining the lethality of SPS mutations in this strain16,17. We confirmed that deletions of these genes were viable in yeast that could synthesize leucine but were lethal in yeast that could not (Fig. 2b). Similarly, the PTC-tolerant gene FUR1 is required for the utilization of exogenous uracil18, and uracil biosynthesis is also disrupted in the strain used to annotate essential genes. We confirmed that FUR1 is essential only in yeast that cannot synthesize uracil (Fig. 2b), in agreement with previous synthetic-lethality results18. Unexpectedly, we also observed poor growth of yeast with deletions of both URA3 and the PTC-tolerant gene SDH3 (Fig. 2b), a member of the mitochondrial inner-membrane protein translocase complex19, thus suggesting that proper uracil utilization may involve an unknown mitochondrial function. These examples illustrate that genes not universally essential for yeast viability can appear essential in a specific genetic background. Another PTC-tolerant gene, MMF1, was viable in our growth conditions but not in those used to define the set of essential genes (Supplementary Fig. 9 and Supplementary Note 1), thus providing an example of environment-dependent essentiality.
Six PTC-tolerant essential genes encode proteins with large dispensable C-terminal regions. One striking case is CWC24, a highly conserved member of the spliceosome. Cwc24 has a CCCH-type zinc-finger domain (Znf) and a RING-type Znf domain. Analysis of the effect of PTCs in CWC24 suggested that the RING-finger domain was dispensable, but the CCCH Znf was essential, results that we confirmed by engineering CWC24 truncations (Fig. 3a and additional information in ref. 31). Interestingly, RNF113A, the essential human homolog of CWC24, can also tolerate a PTC after its RING-finger domain33. Four other PTC-tolerant genes, TAF7, TAF8, COG3, and LSM4, have been reported to tolerate large truncations34–36. We verified that SEC5, a 971–amino acid member of the essential exocyst complex37, tolerates truncation of at least 615 amino acids (Fig. 3b). Our observation that 101 genes tolerated five or more PTCs suggests that many additional genes may have dispensable C-terminal regions.

Our results improve the annotation of essential genes in the well-studied yeast genome. We discovered several cases of genes that appeared to be essential as a consequence of the specific strain and growth conditions originally used to test the viability of gene deletions. These results were consistent with recently reported results based on transposon mutagenesis38 (Supplementary Note 2 and Supplementary Fig. 10). A deletion screen in a different yeast isolate has also highlighted examples of conditionally essential genes39. Applying our approach and related methods in a diverse set of isolates and growth conditions should further refine the core set of essential yeast genes.

PTCs are prioritized in studies of human genetic variants because of their high likelihood of abolishing gene function. Our results suggest that PTCs are most likely to be deleterious when they disrupt annotated protein domains or truncate more than 27 amino acids, and these criteria may improve filtering of candidate causal variants. We observed that NMD did not strongly contribute to PTC tolerance. This result is consistent with recent findings that NMD in yeast acts most strongly on transcripts with PTCs toward their 3′ ends40. PTCs near the ends of human genes are also likely to escape NMD, according to the 50-base-pair rule41 (Supplementary Fig. 11), and our criteria may be especially useful for predicting their effects.

In our study, we carried out a pooled screen of the functional endpoints at eight loci in yeast tetrads, as determined by classifying paired-end Illumina reads of PCRs of genomic DNA at each locus. Outcomes of directed mutations at eight loci in yeast tetrads, as determined by classifying paired-end Illumina reads of PCRs of genomic DNA at each locus.

| Gene          | Expected edit (%) | Unedited (%) | Mismatch (%) | Indel (%) |
|---------------|-------------------|--------------|--------------|-----------|
| ho-G582Stop   | 98.51             | 0.08         | 1.40         | 0.00      |
| his2-E308Stop | 99.83             | 0.07         | 0.10         | 0.00      |
| mnt1-V219Stop | 99.35             | 0.53         | 0.12         | 0.00      |
| spo1-F381Stop | 95.56             | 4.24         | 0.19         | 0.00      |
| spo13-P252Stop| 99.67             | 0.20         | 0.13         | 0.00      |
| ste3-P469Stop | 99.75             | 0.12         | 0.13         | 0.00      |
| can1-G121Stop | 99.81             | 0.06         | 0.13         | 0.00      |
| can1-G70Stop  | 99.80             | 0.03         | 0.17         | 0.00      |

Outcomes of directed mutations at eight loci in yeast tetrads, as determined by classifying paired-end Illumina reads of PCRs of genomic DNA at each locus.

Fig. 3 | Selected truncatable essential genes. a, Tolerance scores for ten PTCs in CWC24 are shown by gray circles; red and green bars indicate hidden Markov model (HMM) calls of deleterious and tolerated, respectively (top). The RING-finger and CCCH Znf domains of Cwc24 are highlighted. Analysis of deleterious and tolerated truncations of CWC24 in yeast tetrads, displayed as in Fig. 2 (bottom). Deletions of the last 88 and 94 codons of CWC24 are tolerated (middle and right), whereas deletion of the last 119 codons is not (left). n = 10 tetrads were examined for each tested deletion. b, Tolerance scores for eight PTCs in SEC5 are shown by gray circles; red and green bars indicate HMM calls of deleterious and tolerated, respectively (top). The Pfam-annotated SEC5 domain is highlighted. Analysis of deleterious and tolerated truncations of SEC5 in yeast tetrads (bottom). Deletion of the last 615 codons of SEC5 is tolerated (right), whereas deletion of the last 707 codons is not (left). n = 8 tetrads were examined for the deletion of 615 codons, and n = 12 tetrads were examined for the deletion of 707 codons.

These features of our method enable many applications in yeast, including targeted genome-wide mutagenesis screens, deep mutational scanning of specific genes, and assessment of phenotypic effects of natural variants. Notably, the method can be used in strains other than the S288c reference strain, for which tools such as deletion libraries are largely unavailable.

Multiplex CRISPR-based editing has been reported at single loci in human cells42. This method differs from ours in that it uses a single gRNA in combination with a library of repair templates to generate many distinct edits in a small genomic region. The method also requires selective enrichment of edited DNA as a consequence of the low usage of HDR in DSB repair in human cells. This general limitation applies to all uses of editing in mammalian cells, including potential extensions of our method. Another challenge to extending our method from yeast to mammalian cells is the need...
for longer homology regions in the repair templates. Improving DNA synthesis and delivery and enhancing the efficiency of HDR in mammalian cells are active areas of research\textsuperscript{35–37}, and we anticipate that advances on these fronts will facilitate the development of a mammalian version of our system.

The approach described here could be extended to assess the functional effects of any desired nucleotide variants in a highly parallel manner. The ability to profile the effects of broad classes of alleles, including missense and regulatory variants, would enable a more fine-grained understanding of the relationship between genotypes and phenotypes.

URLs. CRISPR efficiency predictor tool, \url{http://www.flyrnai.org/evaluateCrispr/}; Saccharomyces Genome Database, \url{http://yeastgenome.org/}.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at \url{https://doi.org/10.1038/s41588-018-0087-y}.

Received: 14 August 2017; Accepted: 21 February 2018; Published online: 9 April 2018

References

1. Shailem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR-Cas9. \textit{Nat. Rev. Genet.} 16, 299–311 (2015).
2. DiCarlo, J. E. et al. Genome engineering in \textit{Saccharomyces cerevisiae} using CRISPR-Cas systems. \textit{Nucleic Acids Res.} 41, 4336–4343 (2013).
3. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. \textit{Nucleic Acids Res.} 44, D682–D688 (2016).
4. Stenson, P. D. et al. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. \textit{Hum. Genet.} 133, 1–9 (2014).
5. Klasson, H., Fink, G. R. & Ljungdahl, P. O. Ssy1p and Ptr3p are plasma membrane Ssy1p-Ptr3p-Ssy5p sensor of extracellular amino acids. \textit{Genes Dev.} 15, 5405–5416 (1999).
6. Decourty, L. et al. Long open reading frame transcripts escape nonsense-mediated mRNA decay in yeast. \textit{Cell Rep.} 6, 593–598 (2014).
7. Brill, B. et al. Bromodomain factor 1 corresponds to a missing piece of yeast TFIID. \textit{Nat. Biotechnol.} 33, 543–548 (2015).
8. Dorcy, L. et al. Long open reading frame transcripts escape nonsense-mediated mRNA decay in yeast. \textit{Cell Rep.} 6, 593–598 (2014).
9. Nagy, E. & Maquat, L. E. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. \textit{Trends Biochem. Sci.} 23, 196–199 (1998).
10. Garst, A. D. et al. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. \textit{Nat. Biotechnol.} 35, 48–55 (2017).
11. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation editing of genomic regions by multiplex homology-directed repair. \textit{Nature} 513, 120–123 (2014).
12. Chopra, M. et al. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. \textit{Nat. Biotechnol.} 33, 543–548 (2015).
13. Lin, S., Staal, B. T., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. \textit{eLife} 3, e04766 (2014).
14. Song, J. et al. RS-1 enhances CRISPR/Cas9- and TALEN-mediated knock-in efficiency. \textit{Nat. Commun.} 7, 10548 (2016).

Acknowledgements

We thank members of the laboratory of L.K., and F. Albert, M. P. Hughes, and J. Rine for helpful discussions; R. Cheung and E. Pham for technical assistance; and G. Church (Harvard Medical School) for plasmids. Funding was provided by the Howard Hughes Medical Institute and NIH grants R01 GM102308 (L.K.) and F32 GM116318 (M.J.S.).

Author contributions

Experiments were designed by M.J.S., J.S.B., and L.K. Experiments were performed by M.J.S. and L.D. Data were analyzed by M.J.S., J.S.B., and L.K. The manuscript was written by M.J.S., J.S.B., and L.K., and incorporates comments from all other authors.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at \url{https://doi.org/10.1038/s41588-018-0087-y}.

Reprints and permissions information is available at \url{www.nature.com/reprints}.

Correspondence and requests for materials should be addressed to M.J.S. or J.S.B. or L.K.

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
Methods

Strains and plasmids used. Strains used in this study are listed in Supplementary Table 5; plasmids used in this study are listed in Supplementary Table 6; and oligonucleotides used in cloning are listed in Supplementary Table 7. The plasmids used for cloning the targeting-plasmid pool can be obtained from https://www.addgene.org/Leonid_Kruglyakov/.

Assessing the efficiency of edit-directing plasmids. We designed nine edit-directing plasmids to assess the efficiency of the method. The targeted edits were chosen such that we expected minimal fitness effects (Table 1); each targeted a PTC to a gene that was not expected to be important for vegetative yeast growth. The sites were distributed across six chromosomes. We included two gRNAs targeting CAN1 that have previously been characterized, including one (can1-G121Stop) that targeted a site reported to have lower-efficiency Cas9 targeting27. To generate edit-directing plasmids, we synthesized DNA fragments carrying the desired gRNA as well as the appropriate repair template as gBlocks (Integrated DNA Technologies), which were then cloned into pLK78 (p426-5SNR2p-gRNA.CAN1.Y.Sup4t, a plasmid for expressing gRNAs under an SNR52 promoter, kindly provided by G. Church) through Gibson assembly28.

YLK3221 (nef1Δ) with pLK77 was transformed with edit-directing plasmids through standard lithium acetate transformation29. YLK3228 (nef1Δ/Δ) with pLK77; Supplementary Table 8), YLK2525 (nef1Δ NMD2) with pLK77 (Supplementary Table 1), and YLK2357 with pLK77 (Supplementary Table 2) were also transformed with select edit-directing plasmids to test the effects of NMD2, NEJ1, and diploidy, respectively, on editing efficiency. pLK77 encodes Streptococcus pyogenes Cas9 under the Gal1L galactose-inducible promoter and was kindly provided by G. Church2. Edit-directing plasmids were transformed into yeast on galactose plates to induce Cas9 expression. Single colonies were picked and grown in glucose overnight, after which approximately 1,000 cells of each culture were plated on galactose plates to induce Cas9 expression. Colonies were allowed to grow for approximately 72 h, after which cells were washed off the plates. Genomic DNA was extracted from harvested cells with a DNeasy Blood and Tissue kit (Qiagen), with a yeast-optimized protocol using Zymolase (Amresco). For each targeted site, genomic DNA was amplified with Pfu Ultra II (Agilent Technologies). To specifically amplify genomic DNA rather than the edit-directing-plasmid repair template, DNA primers were chosen to anneal outside the regions of homology in the repair templates.

The primer pairs were designed to create PCR products that matched the product of the Illumina Nextera transposon reaction (Illumina). Indexed sequencing libraries were generated from these PCR products by proceeding with amplification and indexing, as described in the Nextera DNA Library Prep Protocol Guide. Libraries were pooled before sequencing on an Illumina MiSeq sequencer with 300-bp paired-end reads. Our PCR products were designed to generate paired-end reads with extensive regions of overlap; these overlaps were stitched together with PEAR, version 0.9.6 (with parameters -v10 -m 400 --q 20), to decrease errors introduced by mistaken sequencing calls. Reads were trimmed to remove adaptor sequences in trimmomatic, version 0.32 (ref. 42), and matched to the set of expected edits. For each oligonucleotide’s frequency of adenine bases. From the oligonucleotide synthesis of the reverse primer. Two separate possible barcode-sequence classes were used, to generate two edit-directing-plasmid pools distinguishable by their barcodes. One pool had barcodes of the form NNNNNNNNNSSWWS, and the other had barcodes of the form NNNNNNNNNSSWS, where S is either a cytosine or guanine base, W is either an adenine or thymine base, and N is any of the four bases. 10 μg of each PCR product was digested with MluI-HF and Sphi-HF and gel extracted with a QIAquick Gel Extraction kit (Qiagen).

Generation of barcoded edit-directing-plasmid pools. The single-stranded oligonucleotides were synthesized on the Oligo Library Synthesis (OLS) platform (Agilent Technologies) in either the Watson or the Crick orientation to minimize each oligonucleotide’s frequency of adenine bases. From the oligonucleotide pool generated by OLS, we generated edit-directing-plasmid pools via a ligation-mediated cloning scheme described below (graphically summarized in Supplementary Fig. 1). Oligonucleotides were amplified on an AriaMax real-time PCR system (Agilent Technologies) with a KAPA Library Amplification kit (Kapa Biosystems). The amplification primers were designed to introduce an EagI cut site into the 3′ end of the amplification product (Supplementary Table 7, primers named OLS Library Amplification F and R). Reactions were stopped during linear amplification, and the amplified library was then purified with a QIAquick PCR purification kit (Qiagen), digested with BstEEII-HF and Eagl-HF (New England Biosciences), and purified again.

The amplified library was cloned into pLK88, a version of pLK78 modified to include BstEEII and Sphi sites. pLK88 was isolated with a Plasmid Plus Maxiprep kit (Qiagen) from 200 ml of Escherichia coli culture. 20 μg of plasmid was then digested with BstEEII-HF and Eagl-HF, treated with shrimp alkaline phosphatase (New England Biolabs), and purified with a Qiagen PCR purification kit. We tested two ligation reactions: 1 μg of digested vector was ligated with either 100 or 800 ng of the digested insert, with 4 μl of T4 DNA ligase M0202M (New England Biolabs), in an 800-μl or 200-μl reaction, respectively, at room temperature for 10 min. Concurrently, we ran negative-control ligations lacking the insert DNA. Ligation reactions were stopped on ice. To test ligation efficiency, 0.5 μl of each ligation was transformed on ONESHOT electro-competent E. coli cells (Lucigen) with a Bio-Rad plasmid mini-prep kit (Bio-Rad Laboratories). After 1 h of rescue growth, cells were transferred to 200 ml of LB medium with 100 μg/ml ampicillin (Sigma-Aldrich) and grown overnight. From serial dilutions plated after the transformation, we estimated that approximately 700,000 E. coli cells were transformed. Plasmids were maxiprepared from 150 ml of culture.

Next, we cloned in the remaining gRNA structural region and terminator between the gRNA targeting sequence and repair template of the first-step cloning product, while adding a 12 nt-bracketed adenine to the repair template (Supplementary Fig. 1). The cloning insert also included a kanamycin-resistance gene to facilitate enrichment of the correct cloning product. The insert sequence was amplified from pLK89 (the amplification primers used are named ‘Insert Amplification’ in Supplementary Table 7) with Plu Ultra II (New England Biolabs) and purified with a Qiagen PCR purification kit. The insert sequence was introduced during the insert PCR amplification by use of mixed bases in the synthesis of the reverse primer. Two separate possible barcode-sequence classes were used, to generate two edit-directing-plasmid pools distinguishable by their barcodes. One pool had barcodes of the form NNNNNNNNNSSSWW, and the other had barcodes of the form NNNNNNNNNSSWS, where S is either an cytosine or guanine base, W is either an adenine or thymine base, and N is any of the four bases. 10 μg of each PCR product was digested with MluI-HF and Sphi-HF and gel extracted with a QIAquick Gel Extraction kit (Qiagen).
ampicillin and 50 μg/ml kanamycin (Thermo Fisher Scientific). Plasmid DNA was extracted as described above. We estimated that approximately 1.5 × 10⁶ E. coli cells were transformed with each barcoded pool.

PTC induction in yeast pools. Strains YLK3221 (nej1∆) and YLK3229 (nej1∆ nmd2∆) were each separately transformed with both the WSSW and SWWS plasmid pools. For these large-scale transformations, cells were grown in yeast extract peptone dextrose medium (YPD) to an OD₆₀₀ of approximately 0.5, and approximately 1 × 10⁶ cells were then used in a transformation reaction with 10 μg of either plasmid pool. Each transformation was plated on four 15-cm YNB + galactose + CSM –Ura –Leu to induce Cas9 (shaking, at 30°C). Then, every 24 h for the next 96 h, we froze pellets of approximately 2 × 10⁶ cells while transforming of cell cultures (approximately 2.2 × 10⁶ to 7 × 10⁶ cells) to 300 ml of medium for continuing growth. During that time, the cultures did not leave the exponential growth phase, because the OD₆₀₀ remained less than 0.75.

PTC repair-template and barcode sequencing. DNA, including edit-directing plasmid DNA, was extracted from the harvested frozen cell pellets with a DNeasy Blood and Tissue kit. We designed PCR primers to delimit the repair-template sequences from the plasmids for Illumina sequencing (the amplification primers used are denoted ‘repair template amplification for Illumina’ in Supplementary Table 7). The PCRs were performed with a KAPA Library Amplification kit, and then sequencing libraries were generated with a unique index for each culture at each time point. Libraries were pooled and sequenced on four lanes of an Illumina HiSeq 2500 sequencing system with 150-bp paired-end index for each culture at each time point. Libraries were pooled and sequenced on

Calculating PTC tolerance scores and gene tolerance scores. To obtain slope (θ) and intercept (α) estimates for each tracked barcoded plasmid, we fit a linear model by using the glm function in R:

\[
\log(E(counts\ time) \over total\ counts) = \alpha + \theta \times \text{time}
\]

in which we normalized for the differing read depths for the distinct time points by including total_counts, the vector of total observed read counts across all barcodes for each time point. The observed distribution of slopes (θetas) was plotted and modeled (Supplementary Fig. 14). We interpreted the model representing persisting and depleted barcoded plasmids and classified barcoded plasmids as persisting or depleted according to whether they had a theta estimate above or below −0.025; this value was chosen by visual inspection to best separate the two modes.

To calculate the PTC tolerance score and the gene tolerance score, we fit logistic mixed-effect models on these binarized barcoded plasmid persistences by using the lme4 R package and the glmer function. This analysis method was chosen because it handles the nested structure of the experiment, with varying numbers of barcodes per targeted PTC and multiple PTCs per gene. The dependent variable was the classification of each barcoded plasmid as persisting or depleted, as determined above. We included the following fixed effects to control for technical factors that might affect CRISPR function: (i) an indicator variable for potential off-target gRNA hits, as determined by BLAST searching, (ii) an indicator variable for having TTTT in the gRNA targeting sequence (potential terminator for Pol III), (iii) a gRNA efficiency score, as calculated with the CRISPR Efficiency Predictor tool from the Drosofila RNAi Screening Center, (iv) the total number of barcodes observed for each PTC, (v) the strain, as estimated in eq. (1), representing an estimate of the number of observations of each barcode at the initial time point, (vi) the GC content of the gRNA sequence, and (vii) the number of bases edited to generate the PTC. These terms were included to control for potential confounding technical effects on PTC tolerance.

Because PTCs behaved similarly in the nei1∆ or nei1∆ nmd2∆ strain backgrounds (Supplementary Fig. 4b), for subsequent analyses and figures we used both datasets in the models and included a fixed effect (vii) for whether the barcoded PTC was observed in the nei1∆ or nei1∆ nmd2∆ strain background (Figs. 2 and 3, and Supplementary Figs. 3 and 5–7). Thus, certain figures do not show results from the nei1∆ strain background only, and Supplementary Fig. 4b (top) shows results from the nei1∆ nmd2∆ strain background only. The models included or included two random effects: one for the specific PTC directed by each barcoded plasmid and optionally one for the gene targeted by each barcoded plasmid. For technical factors that might affect CRISPR function: (i) an indicator variable for having TTTT in the gRNA targeting sequence (potential terminator for Pol III), (ii) an indicator variable for potential off-target gRNA hits, as determined by BLAST searching, (iii) a gRNA efficiency score, as calculated with the CRISPR Efficiency Predictor tool from the Drosofila RNAi Screening Center, (iv) the total number of barcodes observed for each PTC, (v) the strain, as estimated in eq. (1), representing an estimate of the number of observations of each barcode at the initial time point, (vi) the GC content of the gRNA sequence, and (vii) the number of bases edited to generate the PTC. These terms were included to control for potential confounding technical effects on PTC tolerance.

To determine the correlations for both PTC tolerance and gene PTC tolerance scores between replicate experiments, the model described above was fit independently on SWWS- and WSSW-barcoded plasmids from the nei1∆ background (Supplementary Table 11). This analysis indicated that the scores from the replicate experiments were well correlated, and we therefore combined the replicates into one dataset for all subsequent analyses to increase statistical power.

Gene Ontology enrichment analysis was performed with TopGO, version 2.30.0, and is presented in Supplementary Table 4 (ref. 4).

Extended model with additional features. To determine how features of the barcoded PTCs and their targeted genes affected PTC tolerance, we extended the mixed-effects-model analysis. We restricted the analysis to essential genes only. In addition to the eight fixed effects and two random effects listed above, we modeled the following fixed-effect terms (Supplementary Table 3): (ix) whether the PTC directed by an annotated PTC or not (Supplementary Fig. 3). This analysis indicated that the scores from the replicate experiments were well correlated, and we therefore combined the replicates into one dataset for all subsequent analyses to increase statistical power.

Gene Ontology enrichment analysis was performed with TopGO, version 2.30.0, and is presented in Supplementary Table 4 (ref. 4).
categorizations, as determined by Liu et al.17, (iii) the distance in amino acids of the PTC from the C-terminal end of the protein, (xiv) whether the PTC disrupted a low-complexity region15, (xvi) whether any allele of the gene had ever been annotated as viable in the Saccharomyces Genome Database, and (xvi) the overall length of the gene targeted.

Coefficients were obtained as above from the glmer function (Supplementary Table 3). Type III analysis-of-variance tables were computed for the fixed-effect terms in the model with the Anova() function in the car R package, version 2.1–6 (ref.14). Likelihood-ratio chi-square values and P values for the fixed-effect terms in the model were also computed with this function. Tjur’s D was used to calculate a pseudo R\(^2\) statistic for overall model fit (Tjur’s D=0.39) (ref.14).

Segmented regression. We fit two-segment segmented regressions for PTC tolerance scores given the distance of a PTC, in codons, from the 3′ end of a targeted gene and obtained 95% confidence intervals for the breakpoint, using functions provided in the R package segmented, version 0.5–3.0 (ref.15).

Hidden Markov model of dispensable 3′ ends of genes. We built an HMM for each gene to more precisely delineate the dispensable 3′ ends of essential genes. We ordered PTCs from the 3′ to 5′ end of the gene. For each introduced PTC in a gene, the binary hidden states represented whether the PTC was deleterious or tolerated. The observations were the binarized barcoded plasmid persistences of each PTC, as described above in the section ‘Calculating PTC tolerance scores and gene tolerance scores.’ We assumed that all PTCs 5′ of a deleterious PTC would also be deleterious. This assumption was represented in the model by setting the transition probabilities (moving from the 3′ to 5′ end of the gene) from deleterious to deleterious at 1 and from deleterious to tolerated at 0. The transition probabilities from tolerated to deleterious and from tolerated to tolerated were set at 0.5. We also set the prior probabilities of the 3′–most PTC being deleterious or tolerated at 0.5. The depletion or persistence of an individual barcoded plasmid may not always faithfully represent whether the underlying targeted PTC is deleterious or tolerated. For instance, some gRNAs may target with lower efficiency, thus potentially making a deleterious PTC appear tolerated. In contrast, some tolerated PTCs may appear deleterious if the corresponding edit-directing plasmids drop out of the pool because of off-target cutting or stochastic fluctuations in frequency. To take such errors into account, we set the emission probabilities of a tolerated PTC generating an observation of a barcoded plasmid as persisting at 0.5676 and as depleted at 0.4324. These parameters were estimated as the fractions of barcoded plasmids targeting dubious ORFs that were classified as persisting and depleted. Similarly, we set the emission probabilities of a deleterious PTC generating an observation of a barcoded plasmid as depleted at 0.8056 and as persisting at 0.1944. These parameters were estimated as the fraction of barcoded plasmids at the 5′-most position of essential genes that were classified as depleted and persisting. We used the Viterbi algorithm to identify the most likely hidden state for each PTC.

Testing individual genes. Essential genes predicted to be particularly PTC tolerant or PTC intolerant were verified through direct modification of yeast genes, either through ORF deletion or ORF truncation via partial deletion. The partial deletions of targeted genes were designed as tandem deletions in diploid yeast strains, which were then sporulated in liquid medium. In brief, cells were grown to log phase in rich medium (YPD orYP-galactose) to whichYP with 2% potassium acetate was added for overnight growth, then were transferred to H\(_2\)O with 2% potassium acetate and minimal supplementation of nutrients if required by strain auxotrophies. After asc were observed under a light microscope (typically at 3–7 d), asci were digested with Zymolyase, and tetrads were dissected under a Singer MSM 400 System dissection microscope (Singer Instruments). Unless otherwise mentioned, tetrads dissections were done on YPD plates and photographed after 2 d of growth. Dissections done on ‘defined medium’ plates used YNB+CSM. Genotypes of resulting colonies were determined by replica plating.

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

Data availability. Code and data have been deposited in GitHub at https://github.com/joshbloom/coupledCRISPR_essentialStops/. Illumina read sequences have been deposited in the Sequence Read Archive under project accession PRJNA421550.

References

38. Bao, Z. et al. Homology-integrated CRISPR-Cas (HI-CRISPR) system for one-step multigene disruption in Saccharomyces cerevisiae. ACS Synth. Biol. 4, 585–594 (2015).

39. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods 6, 343–345 (2009).

40. Becker, D. M. & Lundblad, V. Introduction of DNA into yeast cells. Curr. Protoc. Mol. Biol. 27, 13.7 (2001).

41. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina paired-end read merger. Bioinformatics 30, 614–620 (2014).

42. Bolger, A. M., Lahse, M. & Usadel, B. Trimomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).

43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760 (2009).

44. Jiang, W., Bukard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. Nat. Biotechnol. 31, 233–239 (2013).

45. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBNet J. 17, 10 (2011).

46. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48 (2015).

47. Seacue, S. R., Casella, G. & McCulloch, C. E. Variance Components (Wiley, Hoboken, NJ, USA, 2006).

48. Nielsen, S., Yuzenkov, Y. & Zenkin, N. Mechanism of eukaryotic RNA polymerase III transcription termination. Science 340, 1577–1580 (2013).

49. Alexe, A. & Rahmanfaher, J. topGO: Enrichment Analysis for Gene Ontology. https://doi.org/10.18129/B9.bioc.topGO (2016).

50. Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A. & Caves, L. S. D. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22, 2695–2696 (2006).

51. Liu, G. et al. Gene essentiality is a quantitative property linked to cellular evolvability. Cell 163, 1388–1399 (2015)

52. Wootton, J. C. & Federhen, S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol. 266, 554–571 (1996).

53. Fox, J. & Weisberg, S. An R Companion to Applied Regression (Sage, Thousand Oaks, CA, USA, 2011).

54. Tjur, T. Coefficients of determination in logistic regression models: a new interpretation of the coefficient of determination. Am. Stat. 63, 366–372 (2009).

55. Muggio, V. M. R. segmented: an R Package to fit regression models with broken-line relationships. R News 8, 20–25 (2008).

56. Curran, K. A. et al. Short synthetic terminators for improved heterologous gene expression in yeast. ACS Synth. Biol. 4, 824–832 (2015).
Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

1. Sample size
   Describe how sample size was determined.
   We designed an experiment targeting more than 10,000 sites in the genome for engineering premature termination codons (PTCs), corresponding to 10 targets per gene, giving us sufficient replication to measure gene-specific PTC tolerances. In addition, edit-directing plasmids were individually barcoded, with around 10 barcodes observed per edit-directing plasmid, giving us sufficient replication to measure site-specific tolerance. In addition, the experiment was performed twice, and replication correlations are noted in the manuscript.

2. Data exclusions
   Describe any data exclusions.
   No data was excluded from the analyses. Various quality control filters were applied during analysis of sequencing data, as described in the methods.

3. Replication
   Describe whether the experimental findings were reliably reproduced.
   Experimental findings were reliably reproduced.

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups.
   The study design did not require random allocation of samples into experimental groups, as the effect of each mutation could be tracked over time in bulk after its introduction.

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   The bulk and parallel design of the experiment ensured that all samples were subject to identical experimental conditions. Analyses were conducted with automated scripts, which ensured uniform treatment of data.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.
6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

| Item | Confirmed |
|------|-----------|
| 1. The exact sample size \( (n) \) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.) | ☑ |
| 2. A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly | ☑ |
| 3. A statement indicating how many times each experiment was replicated | ☑ |
| 4. The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section) | ☑ |
| 5. A description of any assumptions or corrections, such as an adjustment for multiple comparisons | ☑ |
| 6. The test results (e.g. \( P \) values) given as exact values whenever possible and with confidence intervals noted | ☑ |
| 7. A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range) | ☑ |
| 8. Clearly defined error bars | ☑ |

See the web collection on statistics for biologists for further resources and guidance.

7. Software

Describe the software used to analyze the data in this study.

The following software was used: PEAR 0.9.6, trimmomatic 0.32, bwa 0.7.12-5, TopGO 2.30.0, bio3d 2.3-3, car 2.1-6, cutadapt 1.15, Biostrings 2.46, segmented 0.5-3.0. Additional code was executed in R 3.2.5 with custom R scripts, which is deposited at https://github.com/joshsbloom/coupledCRISPR_essentialStops

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

8. Materials and reagents

Plasmids used for generating the mutation-directing plasmid pool have been deposited at addgene.

No antibodies were used.

Only yeast cells were used. Strains are listed, along with source, in Supplementary Table 5.

Not applicable.

Not applicable.
## Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the [ARRIVE guidelines](#).

### 11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No research animals were used.

### 12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The study did not involve human research participants.