Systematic perturbation of yeast essential genes using base editing

Philippe C Després[1,2,3], Alexandre K Dubé[1,2,3,4], Motoaki Seki[5], Nozomu Yachie*[5,6,7] and Christian R Landry*[1,2,3,4]

1. Département de Biochimie, Microbiologie et Bio-informatique, Faculté de sciences et génie, Université Laval, Québec, Québec, G1V 0A6, Canada

2. PROTEO, le regroupement québécois de recherche sur la fonction, l'ingénierie et les applications des protéines, Université Laval, Québec, Québec, G1V 0A6, Canada

3. Centre de Recherche en Données Massives (CRDM), Université Laval, Québec, Québec, G1V 0A6, Canada

4. Département de Biologie, Faculté de sciences et Génie, Université Laval, Québec, Québec, G1V 0A6, Canada

5. Research Center for Advanced Science and Technology, Synthetic Biology Division, University of Tokyo, Tokyo, 4-6-1 Komaba, Meguro-ku, 153-8904, Japan

6. Department of Biological Sciences, Graduate School of Science, the University of Tokyo, Tokyo, Japan

7. Institute for Advanced Biosciences, Keio University, Tsuruoka, Japan

*To whom correspondence should be addressed. CRL: Tel: 1-418-656-3954, Fax 1-418-656-7176, christian.landry@bio.ulaval.ca NY: Tel +81-3-5452-5242 (x55242), Fax +81-3-5452-5241 (x55241), yachie@synbiol.rcast.u-tokyo.ac.jp

Abstract

Base editors derived from CRISPR-Cas9 systems and DNA editing enzymes offer an unprecedented opportunity for the precise modification of genes, but have yet to be used at a genome-scale throughput. Here, we test the ability of an editor based on a cytidine deaminase, the Target-AID base editor, to systematically modify genes genome-wide using the set of yeast essential genes. We tested the effect of mutating around 17,000 individual sites in parallel across more than 1,500 genes in a single experiment. We identified over 1,100 sites at which mutations have a significant impact on fitness. Using previously determined and preferred Target-AID mutational outcomes, we predicted the protein variants caused by each of these gRNAs. We found that gRNAs with significant effects on fitness are enriched in variants predicted to be deleterious by independent methods based on site conservation and predicted protein destabilization. Finally, we identify key features to design effective gRNAs in the context of base editing. Our results show that base editing is a powerful tool to identify key amino acid residues at the scale of proteomes.

Introduction

Recent technical advances have allowed the investigation of the genotype-phenotype map at high resolution by experimentally measuring the effect of all possible nucleotide substitutions in a short DNA sequence. While saturated mutagenesis informs us on the effect of many mutations, it usually covers a single locus or a fraction of it (Fowler and Fields 2014; Gray *et al.* 2018). Because such data is only available at sufficient coverage for a very small number of proteins, general rules on substitution effects must be extrapolated to other, often unrelated proteins. At a lower level of resolution, genome-scale mutations data has mostly been acquired through large-scale loss-of-function strain collections, where the same genetic change (for example, complete gene deletion) is applied to all genes (Winzeler *et al.* 1999; Giaever *et al.* 2002; C. elegans Deletion Mutant Consortium 2012). This approach is a powerful way to isolate each gene's contribution to a phenotype, including fitness, but limits our understanding of the role of specific positions within a locus.

CRISPR-Cas9 based approaches usually cause protein loss of function through indel formation (Shalem *et al.* 2014) or by modifying gene expression levels (Qi *et al.* 2013; Sander and Joung 2014; Smith *et al.* 2016) at many loci in parallel. Again, these approaches generally limit the information gain to one perturbation per locus. There is therefore a strong tradeoff between the resolution of the existing assays and the number of loci or genes investigated. Recent developments in the field now allow for the exploration of the effects of many mutations per gene across the genome. For instance, in yeast, methods for high throughput strain library construction have allowed the measurement of thousands of variant fitness effects in parallel across the genome (Sharon *et al.* 2018; Bao *et al.* 2018; Roy *et al.* 2018). These approaches rely on CRISPR-Cas9 based genome modifications requiring the formation of double-strand breaks followed by repair using donor DNA, which often depends on complex strain and plasmid constructions. An alternative approach would be to use base editors, which allow the introduction

of the mutations of interest directly in the genome by direct modification of DNA bases rather than DNA segment replacement.

Base editors use DNA modifying enzymes fused to modified Cas9 or Cas12 proteins to create specific point mutations in a target genome (Nishida *et al.* 2016; Gaudelli *et al.* 2017; reviewed in Rees and Liu 2018). Such base editors have recently been used to perform site-specific forward mutagenesis in human cell lines. The two main approaches, Targeted AID-mediated mutagenesis (TAM) (Ma *et al.* 2016) and CRISPR-X (Hess *et al.* 2016), target specific regions of the genome where they induce mutations randomly. This generates a library of mutant genotypes that can be competed to find beneficial and deleterious variants under selective pressure. As the relative fitness measurements depend on targeted sequencing of the locus of interest, these approaches are difficult to adapt to high throughput multiplexed screens where tens of thousands of sites can be targeted within the same gRNA libraries.

Here, we present a method that bridges the flexibility of Target-AID mutagenesis and the multiplexing capacities of genome editing depletion screens. By using a base editor with a narrow and well-defined activity window (Nishida *et al.* 2016), we selected gRNAs generating a limited number of predictable edits in yeast essential genes. This allowed us to use gRNAs as a readout for the effect of the mutations, similar to commonly used barcode-sequencing approaches to measure fitness effects.

Results

Large-scale base editing screening

We used Target-AID mutagenesis to simultaneously assess mutational effects at over 17,000 putative sites in the yeast genome. We scanned yeast essential genes for sites amenable to editing by the Target-AID base editor as well as targets with other specific properties, including

intronic sequences (Figure 1A, Figure S1). Because all essential genes have the same fitness effects when deleted (Giaever *et al.* 2002), focusing on these genes allowed to limit the variation in fitness that could be due to the relative importance of individual genes for growth rather than to the importance of specific positions.

To ensure we could predict gRNA mutational outcomes with accuracy, we included in the library only gRNAs with one to two nucleotides with a high probability of being edited based on the known activity window of Target-AID in yeast (Nishida *et al.* 2016). We could then predict mutagenesis outcomes for gRNAs computationally. We took into account that Target-AID is produces both C-to-G and C-to-T mutations in yeast, with a 1.5 to 2 fold preference for C-to-G (Nishida *et al.* 2016; Després *et al.* 2018). We also extended the analysis to include other point mutants at possible secondary editing sites within the activity window (see methods). As such, we could associate most gRNAs targeting protein-coding DNA to a primary C-to-G and C-to-T outcome (C-to-G #1 and C-to-T #1), as well as to possible secondary outcomes if applicable (C-to-G #2 and C-to-T #2). We did not consider gRNAs that did not target between the 0.5th and 75th percentile of the length of annotated genes to limit position biases that could influence the efficiency of stop-codon generating guides (Doench *et al.* 2014; Michel *et al.* 2017).

The gRNA library was cloned into a high-throughput co-selection base editing vector (Després *et al.* 2018). We performed pooled mutagenesis followed by bulk competition (Figure S2) to identify mutations with significant fitness effects. As the relative abundance of each gRNA in the extracted plasmid pool depends on the abundance of the subpopulation of cells bearing these gRNAs, any fitness effect caused by the mutation they induce will influence their relative abundance. Variation in plasmid abundance was measured using targeted next-generation sequencing of the variable gRNA locus on the base editing vector in a manner similar to GeCKO approaches (Sanjana *et al.* 2014; Shalem *et al.* 2014).

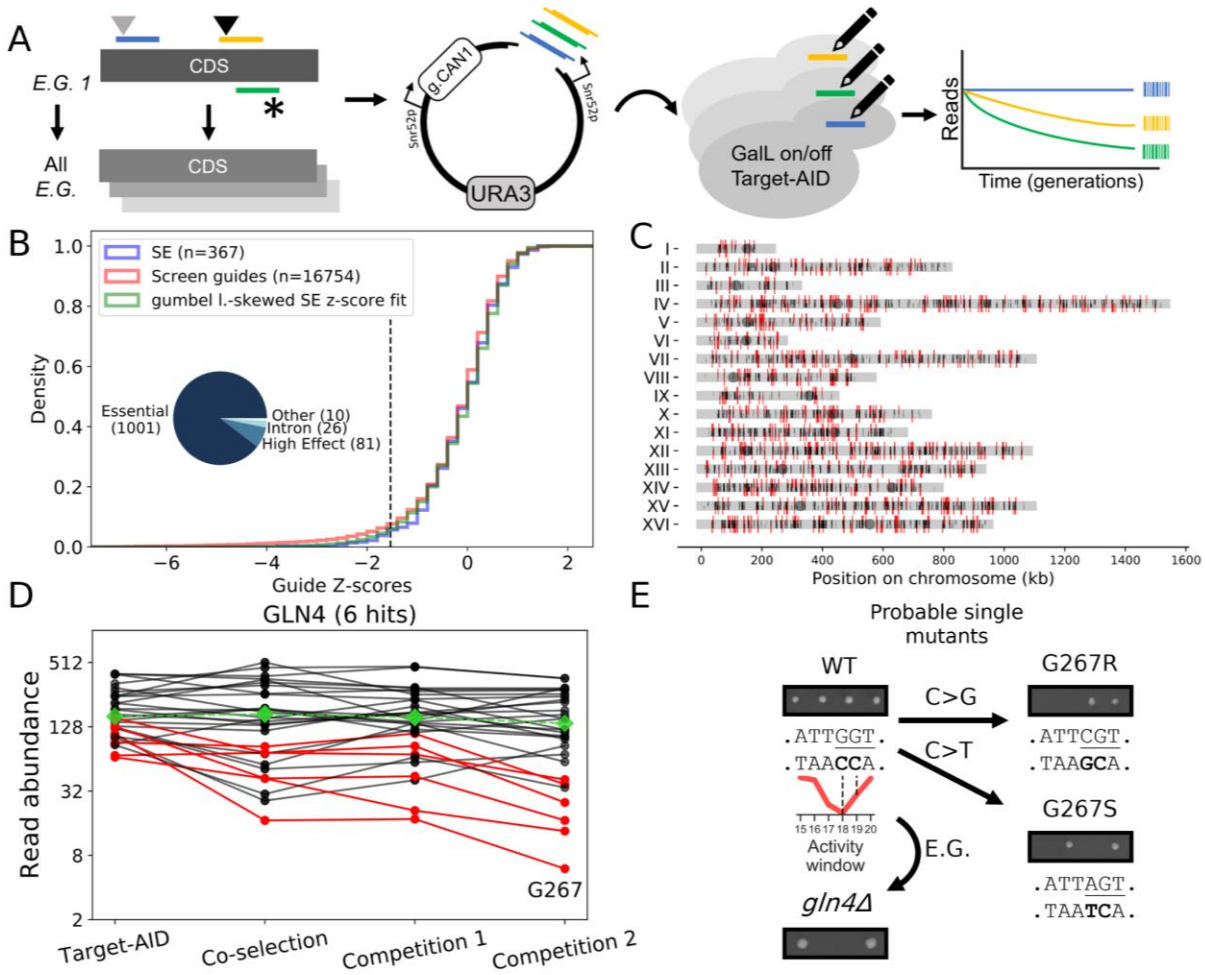


Figure 1. High-throughput forward mutagenesis by Target-AID base editing identifies sensitive sites across the yeast genome. A) Experimental design. Essential genes were scanned for sites appropriate for Target-AID mutagenesis. Mutational outcomes include silent (grey triangle) and missense (black triangle) mutations, as well as stop codons (*). DNA fragments bearing the gRNA sequences were synthesized as an oligonucleotide pool and cloned into a co-selection base editing vector. Using gRNAs as molecular barcodes, the abundance of cell subpopulations bearing mutations was measured during mutagenesis and bulk competition. Mutations with fitness effects were inferred from a reduction in the relative barcode read count. **B)** Cumulative distribution of z-scores of the $\log_2$ fold-change in gRNA abundance between mutagenesis and the end of the bulk competition experiment averaged between replicates (see Figure S2). A 5% false positive threshold was calculated by fitting a distribution of abundance variation z-score of the sequenced gRNAs with synthesis errors (SE gRNAs) and is represented by a dotted black line. The distribution of target types in the 1,118 gRNAs with Negative Effects (GNE) is shown in the inset. **C)** Positions of base editing target sites in the yeast genome. Telomeric regions are depleted in target sites because very few essential genes are located there. GNEs are shown in red, and other gRNAs are in black. The orientation of the line matches the targeted strand relative to the annotated coding sequence. **D)** Decline in barcode abundance (on a log scale) between timepoints after mutagenesis for gRNAs targeting *GLN4*, a tRNA synthetase. Median barcode abundance across the entire library through time is shown in green. The red lines represent the gRNAs categorized as having a significant effect (GNE) for this gene, while non-significant gRNAs (NSG) are shown in black. The gRNA with the most extreme z-score targets residue G267. **E)** Mutagenesis of *GLN4*-G267 confirms its essential role for protein

function (See methods and Figure S3A). Tetrad dissection of a heterozygous deletion mutant bearing an empty vector results in only two viable spores, while the wild-type copy in the same vector restores growth. Dissection of the two heterozygous mutants bearing a plasmid with the most probable single mutant based on the known activity window of Target-AID shows both mutations are lethal.

After applying a stringent filtering threshold based on barcode read count at the mutagenesis step (Figure S2), we identified a total of ~17,000 gRNAs for which we could evaluate fitness effects. Replicate data for gRNAs passing the minimal read count selection criteria show high correlation across experimental time points (Figure S3) and cluster by experimental step (Figure S4), showing that the approach is reproducible. Using the distribution of abundance variation of non-functional gRNAs with synthesis errors as a null distribution (see methods), we identified 1,118 gRNAs across 605 genes or loci with significant negative effects (GNE) on cell survival or proliferation using at an estimated 5% false positive rate. GNEs are distributed evenly across the yeast genome (Figure 1B and 1C), suggesting no inherent bias against specific regions.

An example of barcode abundance variation through time for all gRNAs (both GNEs and NSGs) targeting *GLN4* is shown in Figure 1D. *GLN4* is an essential gene coding for a glutamine t-RNA synthetase. To confirm the deleteriousness of the predicted mutations, we transformed a centromeric plasmid bearing a wild-type or mutated copy of the gene under the control of its native promoter (Ho *et al.* 2009) in a heterozygous deletion background (Giaever *et al.* 1999). Following dissection, spore survival was compared between wild-type and mutated copy of *GLN4* (Figure S5). Using this approach, we confirmed the strong fitness effect of the best scoring GNE for *GLN4*, as the most probable mutations generated are in fact lethal (Figure 1D).

Comparison of GNE induced mutations with variant effect predictions

If GNEs indeed induce specific deleterious mutations, these mutations should be predicted to be more deleterious than those of Non-Significant gRNAs (NSG). We tested two recently published resources for variant effect prediction: Envision (Gray *et al.* 2018) and Mutfunc (Wagih *et al.* 2018). Envision is based on a machine learning approach that leverages large-scale saturated

mutagenesis data of multiple proteins to perform quantitative predictions of missense mutation effects on protein function. The lower the Envision score, the higher the effect on protein function. Mutfunc aggregates multiple types of information such as residue conservation through the use of SIFT (Ng and Henikoff 2003) as well as structural constraints to provide a binary prediction of variant effect based on multiple quantitative and qualitative values. Mutations with a low SIFT score have a lower chance of being tolerated, while those with a positive ΔΔG are predicted to destabilize protein structure or interactions. Both Envision and the Mutfunc aggregated SIFT data cover the majority of the most probable mutations generated by the gRNA library (Figure S6A). The structural modeling information had much lower coverage, covering at best around 12% of the most probable mutations (Figure S6B).

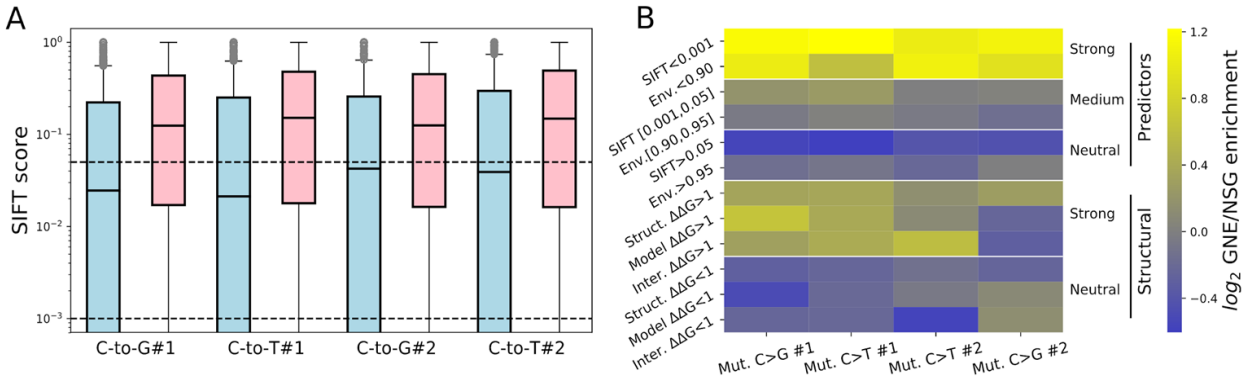


Figure 2: GNE induced mutations are enriched in predicted deleterious effects A) SIFT score distributions for the most likely induced mutations of both GNEs (blue) and NSGs (red). The thresholds for the categories used in the enrichment calculations in **B)** are shown as black dotted lines. SIFT scores represent the probability of a specific mutation being tolerated based on evolutionary information: the first threshold of 0.05 was set by the authors in the original manuscript (Ng and Henikoff 2003) but might be permissive considering the number of mutations tested in our experiment. All GNE vs NSG score comparisons are significant (Welch's t-test p-values: 1.19×10^{-24}, 3.01×10^{-24}, 9.00×10^{-12}, 1.55×10^{-12}). The box cutoff is due to the large fraction of mutations for which the SIFT score is 0. B) Enrichment folds of GNEs over NSGs for different variant effect prediction measurements. Envision score (Env.), SIFT score (SIFT), protein folding stability based on solved protein structures (Struct. ΔΔG), protein folding based on homology models (Model ΔΔG) and protein-protein interaction interface stability based on structure data (Inter. ΔΔG). The raw values used to calculate ratios are shown in Supplementary table 1. The prediction based on conservation and experimental data are grouped under 'Predictors' and those based on the computational analysis of protein structures and complexes under 'Structural'.

As expected, mutations generated by GNEs showed significantly lower SIFT scores (Figure 2A) and showed enrichment for strong effects predicted by SIFT, and Envision. Indeed, all four most probable substitutions created by GNEs are about twice more likely to be predicted to have a large deleterious effect by Envision or a very low chance of being tolerated as predicted by SIFT compared to NSG gRNAs. The high homogeneity of Envision scores across the proteome makes it harder to interpret. As such, the shift in score values is more subtle but supports that GNE mutations are generally more likely to be deleterious as well (Figure S6C, Figure S7A).

Mutation with destabilizing effects as predicted by structural data also appeared to be enriched for the most probable mutations but low residue coverage limits the strength of this association. This is supported by the raw $\Delta\Delta$G value distributions, which show a significant tendency (Welch's t-test p-values: 0.0001, 0.0064, 0.148, 0.007) for GNE mutations to be more destabilizing (Figure S7B,C,D). However, the shift in distribution only achieved significance for certain mutation predictions based on solved structures and homology models. While low residue coverage limits our statistical power, this weak apparent enrichment for mutations affecting protein stability may reflect the marginal stability of the target proteins (DePristo *et al.* 2005), resulting in individual destabilizing mutations having a limited effects on fitness. As expected from known experimental data on mutagenesis outcomes (Nishida *et al.* 2016), signal was usually stronger for the most probable C to G mutation.

Sensitive sites provide new biological insights

Because our screen specifically targeted essential genes, many gRNAs cause mutations in highly conserved regions with high functional importance. To illustrate this, we focus on the highest scoring GNE targeting *GLN4*, a tRNA synthetase, shown in Figure 1D. The gRNA 33725 mutates a glycine at position 267 into either an arginine or a serine. Glycine 267 is part of the "HIGH" motif, characteristic of class I tRNA synthetases, and is involved in ATP binding and catalysis and is highly conserved through evolution (Eriani *et al.* 1990). As expected, the region around the "HIGH"

motif shows both a low evolutionary rate based on inter-species comparisons and a much lower variant density in yeast populations compared to other domains of Gln4 (Figure S3B), showing conservation both on a short and long timescales. Surprisingly, mutagenesis experiments in the bacterial homolog MetRS concluded that mutating this residue from glycine to alanine did not alter significantly catalysis while mutating it to proline had a strong disruptive effect (Schmitt *et al.* 1995). We found that mutating Gly 267 either to Arg and Ser was enough to cause protein loss of function (Figure 1D). Other sensitive sites identified in *GLN4* by our screen are also clustered in regions with slow evolutionary rates. Interestingly, one of these mutations affects residue R568, which has been hypothesized to play a conserved role from bacteria to yeast in the anti-codon and glutamine recognition process (Grant *et al.* 2013).

Since Target-AID can only generate a limited range of amino acid substitutions from a specific coding sequence, we investigated whether any of these mutational patterns were enriched in GNEs (Figure 3A, source data in Supplementary tables S2, S3, and S4). We found several deviations from random expectations in both C-to-G and C-to-T mutation ratios as well as in mutation combination ratios. Three out of four of the mutation pair patterns involving glycine were enriched in GNEs. For example, the Glycine to Arginine or Serine substitutions (as exemplified by guide 33725 targeting *GLN4*) is the second most enriched pattern, being almost four-fold overrepresented in GNE outcomes. This pattern is consistent with the fact that Arginine has properties highly dissimilar to those of Glycine (Sneath 1966), making these substitutions highly deleterious. Furthermore, as Glycine residues are often important components of cofactor binding motifs (eg.: Phosphates) (Copley and Barton 1994) this observation might reflect a tendency for GNEs to alter these sites. Interestingly, genes for which more than one GNE were detected were enriched for molecular function terms linked to cofactor binding (Supplementary table 5). This suggests that the GNEs might indeed have a tendency to affect protein function through mechanisms other than protein or interaction interface destabilization. These protein properties

236 depend on many residues, making them more robust to single amino acid substitutions, whereas
237 cofactor binding may depend specifically on a handful of residues, making these sites critical for
238 function.

239 As expected, there is a strong enrichment for patterns that result in mutation to stop codons: both
240 C-to-G patterns (Tyrosine to stop and Serine to stop) but only one C-to-T pattern (Tryptophan to
241 stop) was overrepresented significantly. Substitutions to stop codon in one outcome also drove
242 enrichment in the other: for example, the link between Serine to Stop (C-to-G) appears to be the
243 cause of the Serine to Leucine (C-to-T) overrepresentation. Both mutation pairs involving mutating
244 a Tryptophan to a stop via a C-to-T mutation: this is not surprising, as the alternative mutations
245 Tryptophan to Serine or Cysteine are also highly disruptive (Sneath 1966). Changes between
246 similar amino acids, which are expected to be tolerable, were also generally depleted in GNE (ex.:
247 the Alanine to Glycine/Valine pair). Mutations in intronic sequences and putative non-functional
248 peptides were also underrepresented, as were most patterns leading to silent mutations. These
249 results show the power of this approach to discriminate important functional sites from mre
250 mutation tolerative ones across the genome.

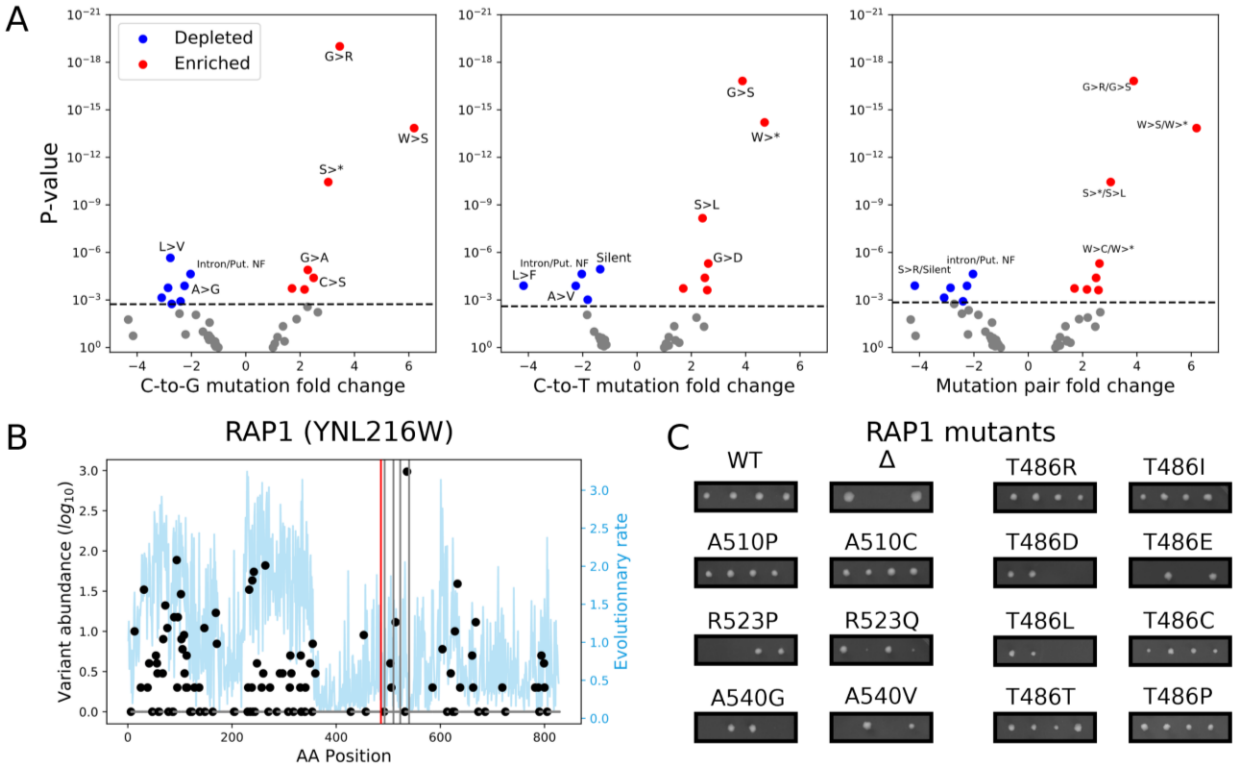


Figure 3 GNE mutations are enriched for specific amino acid substitution patterns and identify critical sites for protein function. A) Fold depletion and enrichment volcano plots for the most probable mutations induced by GNEs in the screen. Enrichment and depletion values were calculated by comparing the relative abundance of each mutation among GNEs and NSGs using Fisher's exact tests. Mutation patterns significantly depleted are shown in blue, while those that are enriched are in red. The significance threshold was set using the Holm-Bonferroni method at 5% FDR and is shown as a dotted grey line. **B)** Protein variant frequency among 1000 yeast isolates (black dots) and residue evolutionary rate across species (blue line) for *RAP1*. The target site for the GNEs targeting T486 is highlighted by a red line while the other detected GNEs target sites are shown by a grey line. **C)** Tetrad dissections confirm most *RAP1* GNE induced mutations indeed have strong fitness effects, as well as other substitutions targeting these sites.

The precise targeting of our method also allows us to investigate amino acid residues with known functional annotations such as post-translational modifications. We found no significant enrichment for gRNAs mutating directly annotated PTMs (ratio$^{\text{GNE PTM}}$ = 19/1118, ratio$^{\text{NSG PTM}}$ 243/15536, Fisher's exact test p=0.71). This is consistent with the hypothesis that many PTM sites may have little functional importance (Landry *et al.* 2009) and thus their mutations may have no detectable effects for a large part. The same was also observed for gRNAs mutating residues near known PTMs that could disturb recognition sites (ratio$^{\text{GNE nearPTM}}$ = 130/1118, ratio$^{\text{NSG nearPTM}}$

= 1698/15536, Fisher's exact test p=0.43). However, GNEs that do target annotated PTM sites might provide additional evidence supporting the importance of these sites in particular. For example, the best scoring GNE in the well-studied transcriptional regulator *RAP1* is predicted to mutate residue T486. This threonine has been reported as phosphorylated in two previous studies (Albuquerque *et al.* 2008; Holt *et al.* 2009), but the functional importance of this phosphorylation has not been explored yet. Residue T486 is located in a disordered region in the DNA binding domains (Konig *et al.* 1996), which part of the only *RAP1* fragment essential for cell growth (Graham *et al.* 1999; Wu *et al.* 2018).

Because the available wild-type *RAP1* plasmid (see methods) does not complement gene deletion growth phenotype, we used a different strategy for confirmation that relied on CRISPR-mediated knock-in (see methods and Figure S8). While we could not confirm that the two most likely mutations predicted to be caused by the GNE had a detectable fitness effect in these conditions, we found that phosphomimetic mutations at this position were lethal (Figure 3C and D) but most other amino acids were well tolerated. This suggests that the constitutive phosphorylation of this residue would be highly deleterious. We could also confirm deleterious effects for GNE induced mutations targeting residues R523 and A540, while mutations at residue A510 had no detectable effect on fitness (Figure 3C and D). As we only tested progeny survival on rich media and at a permissive temperature and the screen was performed in synthetic media at 30°C, these mutants might still affect cell phenotype but in an environment-dependent manner.

gRNA properties influence mutagenesis efficiency

There are still very few high-throughput experimental datasets available that allow the investigation of which gRNA properties affect editing efficiency in the context of base editing. Such large-scale data was key in developing models to optimize Cas9 nuclease activity in other types of genome editing experiments, which revealed that sequence specific motifs and thermodynamic RNA properties can be key features (Doench *et al.* 2014, 2016; Wong *et al.* 2015). As gRNAs

296 showing high Cas9 nuclease activity might have poor base editing activity (Kim *et al.* 2017),
297 existing datasets are not easily transferable. We therefore examine what gRNA and target
298 sequence features could influence mutagenesis efficiency. To do so, we focused on the subset
299 of gRNAs with the potential to generate stop codons (stop codon generating gRNAs, SGG) in
300 essential genes (Figure 4A). Successful mutagenesis by SGGs should result in cell death or no
301 proliferation, and a sharp decrease in read abundance, also serving as a positive control for
302 fitness effects within the screen. As most gRNAs were designed to target the first 75% of the
303 coding sequences of essential genes, it is expected that stop codons in these genes would lead
304 to a loss of function.

305

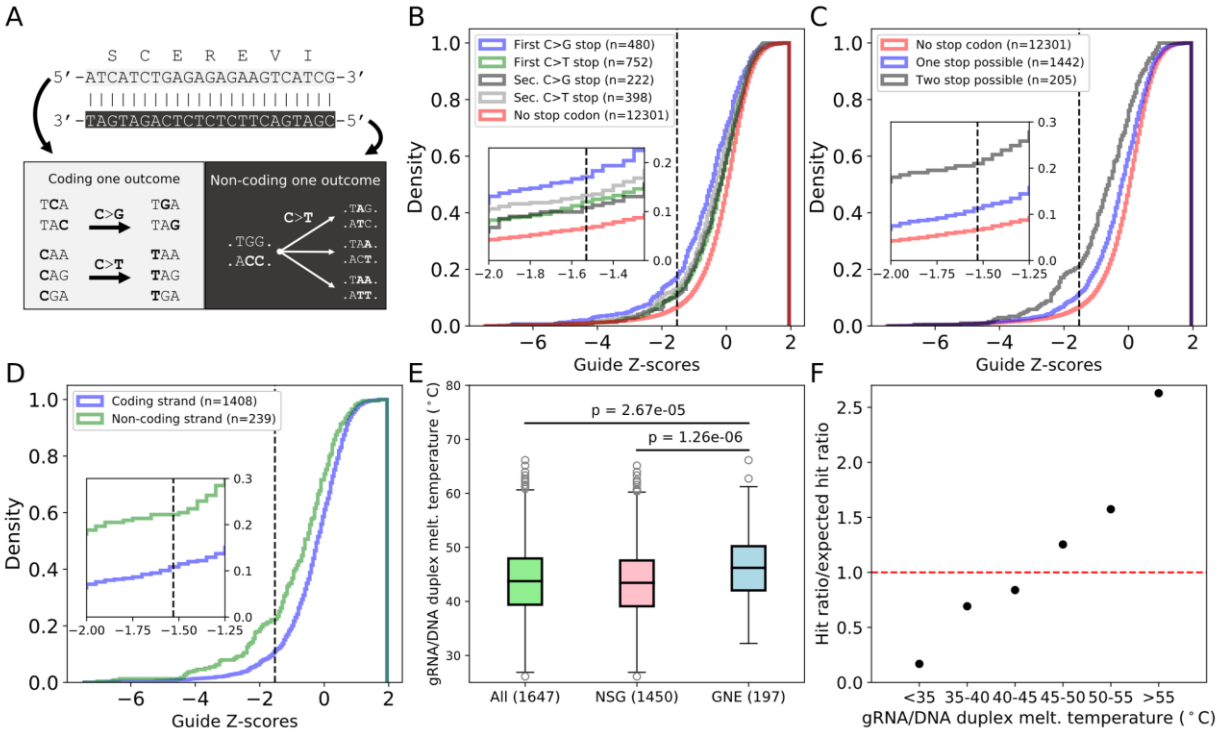


Figure 4: gRNA and target properties affect mutagenesis efficiency. A) Since Target-AID can generate both C to G and C to T mutations, many codons can be targeted to create premature stop codons. **B)** Cumulative z-score density of SGGs grouped by the mutational outcome generating the stop codon. A higher rate of GNE is observed for gRNAs for which a C-to-G mutation at the highest editing activity position generates a stop codon mutation. The significance threshold is shown as a black dotted line **C)** Cumulative z-score density of SGG with a different number of mutational outcomes that could result in a stop codon. gRNAs for which more than one mutational outcome results in a stop codon show a higher mutagenesis success rate. **D)** Cumulative z-score density of SGG targeting the coding or non-coding strand. Stop codon generating gRNAs targeting the non-coding strand of essential genes show higher efficiencies compared to those targeting the coding strand. **E)** Distributions of modeled RNA/DNA duplex melting temperature for all stop codon generating gRNAs, those with no significant effects, and SGG GNE. P-values were calculated using the two-sample Kolmogorov-Smirnov test. **F)** Stop codon generating gRNAs GNE enrichment compared to the expected GNE ratio for different melting temperature ranges.

Data from the original Target-AID study (Nishida *et al.* 2016) suggests that the most prevalent outcome for an edited site is a C-to-G transversion. Our data support this observation, as gRNAs which would lead to a C-to-G mutation at the highest activity site of the editing window have the highest GNE detection rate (Figure 4B). It was also suggested that Target-AID could modify multiple nucleotides within the activity window that could be edited during mutagenesis. Our data support this observation, as gRNAs for which two outcomes have the potential to generate a stop codon are markedly more efficient than those with only one stop codon outcome (Figure 4C). This finding also extends to gRNAs that do not generate stop codons (Figure S9A).

We observed that the targeted strand relative to transcription greatly influenced editing efficiency (Figure 4D). This strand effect can be explained by multiple factors. First, there are multiple outcomes leading to mutation to a stop codon starting from a TGG codon (shown in Figure 4A). This codon is the only one that can be targeted on the non-coding strand to generate a stop codon. Second, repair efficiency has been shown to be higher for the transcribed strand in yeast (Reis *et al.* 2012). Finally, as the non-coding strand is the one which is transcribed, a deamination event there might lead to consequences at the protein level more rapidly because it does not need DNA replication to be present on both strands. gRNAs that do not generate stop codons also have a higher chance of having a fitness effect if they target the non-coding strand (Figure S9B), but we did not observe any effects of the chromosomal strand on efficiency (Figure S9C).

One other parameter with a high impact on mutagenesis rate is the predicted melting temperature of the RNA-DNA duplex formed by the gRNA sequence and its target DNA sequence (Figure 4E). The distribution of the melting temperature shows a clear shift between stop codon generating gRNAs that have an effect on fitness and those that do not. gRNAs with low values have a lower chance of being detected as having effects, while gRNAs with higher values are enriched for GNE (Figure 4F). This observation also extends to gRNAs that do not generate stop codons (Figure S9D, E). This enrichment cannot be attributed to technical biases in library preparation or high-throughput sequencing that would tend to lower their abundance as melting temperature shows practically no correlation with read count at every time point (Figure S10). Furthermore, this effect is not caused by target position bias within target genes or a strong correlation between GC content and the targeted position (Figure S11). As binding energy can differ drastically even within groups of gRNAs with similar GC content (Figure S9F), this could provide a useful criterion to help select efficient gRNAs.

Discussion

We tested whether the Target-AID base editor is amenable for genome-wide mutagenesis. Using the yeast essential genes as test cases, we identified hundreds of gRNAs targeting residues with significant effects on cellular fitness when mutated. The precision and traceability of Target-AID genome editing allowed us to predict the mutational outcomes of GNE and to confirm their effects using orthogonal approaches. We used this data to investigate which factors influence base editing efficiency and found multiple gRNAs and target properties that affect mutagenesis and that could be optimized for future experiments for specific genomic space. By focusing on a few highly relevant variants, we highlighted the power of our approach to generate new biological insights.

In previously published methods such as TAM and CRISPR-X (Hess *et al.* 2016; Ma *et al.* 2016), the semi-random nature of the editing forces the use of mutant allele frequencies as a readout for mutational fitness effects, potentially limiting the scale of the experiments because only one genomic region can be targeted at a time. To complement these approaches, we use more predictable base editing to increase dramatically the number of target loci, albeit at the cost of a lower mutational density. Our results demonstrate the feasibility of base editing screening at a large scale with applications beyond stop codon generation, and future developments will further enhance it. For instance, the use of a base editor with multiple possible mutagenesis outcomes complexifies the prediction of editing outcomes, which can, in turn, make GNE confirmation challenging. Using a base editor that channels mutational outcomes such as cytidine deaminase-uracil glycosylase inhibitor (UGI) fusion can address this problem but decreases the number of mutations explored during the experiment. However, recently published data on cytidine deaminase-UGI fusion has shown they could lead to off-target editing in vivo at a much higher rate compared to adenine base editors or the Cas9 nuclease (Jin *et al.* 2019; Zuo *et al.* 2019). Although there is currently no high throughput data on the off-target activity of Target-AID, data

generated in yeast in the original publication suggests far lower rates than those recently reported in mammalian cells (Nishida *et al.* 2016).

We provide key empirical data on parameters that can be used to optimize base editing efficiency, based on gRNA dependent properties such as target strand and GC content. The results we observed differ from what has been reported for Cas9-based genome editing, in which high gRNA RNA/DNA duplex binding has been associated with lower mutagenesis efficiency (Wong *et al.* 2015). Our data thus confirms the observation that parameters associated with Cas9 editing cannot readily be transferred to base editors (Kim *et al.* 2017). Furthermore, the temperature at which experiments are performed might affect efficiency for certain gRNAs with low gRNA-DNA duplex binding energy and should be considered when designing base editing experiments in different organisms. However, it remains to be confirmed whether the enrichment for certain gRNA properties we observed are specific to Target-AID or will also be transferable to other base editors as this may depend on the enzymatic properties of these proteins.

The field of base editing is rapidly evolving, with new tools being developed constantly. One of the most recent additions to this fast-growing toolkit is engineered Cas9 enzymes with broadened PAM specificities (Nishimasu *et al.* 2018), which have already been shown to be compatible with base editors. More flexible PAM requirements are especially useful for base editing applications, as they increase the number of sites to be edited and also the number of potential gRNAs per site, increasing the chances of choosing optimal properties and thus greater efficiency (Dandage *et al.* 2019). Our method allows an experimental scale which bridges saturation mutagenesis methods and genome-wide knock-out studies, alleviating the current trade-off between mutational diversity and the number of targets genes to generate new biological insights

Methods

Generation of a gRNA library for Target-AID mutagenesis of essential genes in yeast

The Target-AID base editor has an activity window between base 15 to 20 in the gRNA sequence starting from the PAM, and the efficiency at these different positions was characterized in Nishida *et al.* 2016. This allowed us to predict the mutational outcomes for a specific gRNA provided the number of editable bases in the window is not too high. To select gRNAs, we parsed a database of gRNA targets for the *S. cerevisiae* reference genome sequences (strain S288c) (Dicarlo *et al.* 2013) and applied several selection criteria. Since the screen was to be performed in the BY4741 strain, all gRNAs (unique seed sequence, no NAG site) within the database were aligned to the reference genome of that strain using Bowtie (Langmead *et al.* 2009). Only gRNAs with a single perfect alignment were kept for subsequent steps. To select gRNAs amenable to Target-AID base editing, we selected gRNAs with cytosines within the highest activity window of the editor (positions -17 to -19 starting from the PAM). To limit the total number of possible mutational outcomes, gRNAs with three cytosines within the window were removed as well as those with two cytosines at the highest activity positions. Next, we filtered out any gRNA containing a BsaI restriction site to prevent errors during the library cloning step.

The list of essential genes (n=1156) (Winzeler *et al.* 1999; Giaever *et al.* 2002) was used to discriminate between gRNAs targeting essential or non-essential genes (retrieved from http://www-sequence.stanford.edu/group/yeast_deletion_project/Essential_ORFs.txt). Among non-essential genes, data from Qian et al. 2012 (Qian *et al.* 2012) was used to create categories of fitness effects. If the fitness score (averaged across media and replicates) of a gene was below 0.75, it was categorized as “high effect” on fitness. We excluded auxotrophic marker genes as well as *CAN1*, *LYP1*, and *FCY1* because those could be used as co-selection markers (Després *et al.* 2018). Gene deletions with an averaged fitness score between 0.999 and 1.001 were categorized as having “no detectable effect” on fitness. We selected gRNAs targeting essential

and high effect genes, as well as gRNAs targeting a set of 38 randomly chosen no effect genes. To further limit the space of gRNAs examined, only gRNAs mapping from the 0.5th percent to the 75th percent of coding sequences were chosen. We also added gRNAs targeting all known yeast introns (Ares lab Database 4.3) (Grate and Ares 2002) and putative non-functional peptides (Smith *et al.* 2014) selected with the same strategy except for the constraints on gRNA position within the sequence of interest. This resulted in a set of 39,989 gRNAs: library properties are summarized in Figure S1.

Library construction

The plasmids, oligonucleotides, and media used in this study are presented as supplementary tables S6, S7 and S8 respectively. The oligo pool was synthesized by Arbor Biosciences (Michigan, USA) and was cloned into the pDYSCKO vector using Golden Gate Assembly (New England Biolabs, Massachusetts, USA) with the following reaction parameters:

NEB GG buffer 10X	2 µl
pDYSCKO [75ng/ul]	1 µl
Oligo pool [2ng/ul]	1 µl
NEB GG mix	1 µl
Water	15 µl

The ligation mix was transformed in *E. coli* strain MC1061 (*[araD139]$_{B/r}$ Δ(araA-leu)7697 ΔlacX74 galK16 galE15(GalS) λ- e14- mcrA0 relA1 rpsL150(strR) spoT1 mcrB1 hsdR2*, Casadaban and Cohen 1980) using a standard chemical transformation protocol and plated on ampicillin selective media to select for transformants. Serial dilution of cells after outgrowth were plated and then used to calculate the total number of clones produced by the cloning reaction. Quality control of the assembly was performed by Sanger sequencing ~10 clones per assembly reaction. Cells were scraped from plates by adding ~5 ml of sterile water, incubating a few minutes at room temperature, and then using a glass rake to resuspend colonies. Resuspended plates were then

pooled together in a single flask per reaction, which was then used to make glycerol stocks of the library and cell pellets for plasmid extraction. The Qiagen Midi-Prep kit (Qiagen, Germany) was used to extract plasmid DNA from cell pellets by following the manufacturer's instructions. The DNA concentration of each eluate was then measured using a NanoDrop (Thermofisher, Massachusetts, USA), and a normalized master library for yeast transformation was assembled by combining equal quantities of each assembly pool.

Library transformation in yeast

Competent BY4741 (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) cells were first transformed with the pKN1252 (p415-GalL-Target-AID) plasmid using a standard lithium acetate method (Gietz and Schiestl 2007). Transformants were selected by plating cells on SC-L. After 48 h of growth, multiples colonies were used to inoculate a starter liquid culture for competent cells preparation using the standard lithium acetate protocol (Gietz and Schiestl 2007): a culture volume of 200 ml was used to generate enough competent cells for mass transformation. The large-scale library transformation was performed by combining 40 transformation reactions performed with 40 ul of competent cells and 5 ul plasmid library (240 ng/ul) after the outgrowth stage and plating 100 ul aliquots on SC-UL: cells were then allowed to grow at 30°C for 48 h. A 1/1000 serial dilution of the cell recovery was plated in 5 replicates and used to calculate the number of transformants obtained. The total number of transformants reached 3.48 $x10^6$ CFU, corresponding to about 100X coverage of the plasmid pool.

Target-AID mutagenesis and competition screening

The mutagenesis protocol is an upscaled version of our previously published method and is shown in Figure S2. Transformants were scraped by spreading 5 ml sterile water on plates and then resuspending cells using a glass rake. All plates were pooled together in the same flask, and the OD of the yeast resuspension was measured using a Tecan Infinite F200 plate reader (Tecan, Switzerland). Pellets corresponding to about 6 x 10^8 cells were washed twice with SC-UL without

a carbon source and then used to inoculate a 100 ml SC-UL +2% glucose culture at 0.6 OD two times to generate replicates A and B. Cells were allowed to grow for 8 hours before 1 x 10^9 cells were pelleted and used to inoculate a 100 ml SC-UL + 5% glycerol culture. After 24 hours, 5 x 10^8 cells were pelleted and either put in SC-UL + 5% galactose for mutagenesis or SC-UL + 5% glucose for a mock induction control. Target-AID expression (from pKN1252) was induced for 12 hours before 1 x 10^8 cells were pelleted and used to inoculate a canavanine (50 µg/ml) co-selection culture in SC-ULR. After 16 hours of incubation, 5 x 10^7 cells of each culture were used to inoculate 100 ml SC-UR, which was grown for 12 hours before 5 x 10^7 cells were used to inoculate a final 100 ml SC-UR culture which was grown for another 12 hours. Cell pellets were washed with sterile water between each step, and all incubation occurred at 30°C with agitation. ~2 x 10^7 cells were taken for plasmid DNA extraction at the end of each mutagenesis and competition screening step.

Yeast plasmid DNA extraction

Yeast plasmid DNA was extracted using the ChargeSwitch Plasmid Yeast Mini Kit (Invitrogen, California, USA) by following the manufacturer's protocol with minor modifications: Zymolase 4000 U/ml (Zymo Research, California, USA) was used instead of lyticase, and cells were incubated for 1 hour at room temperature, one min at -80°C, and then incubated for another 15 minutes at room temperature before the lysis step. Plasmid DNA was eluted in 70 µl of E5 buffer (10 mM Tris-HCl, pH 8.5) and stored at -20°C for use in library preparation.

Next-generation library sequencing preparation

Libraries were prepared by using two PCR amplification steps, one to amplify the gRNA region of the pDSYCKO plasmid pool and the second to add sample barcodes as well as the Illumina p5 and p7 sequences (Yachie *et al.* 2016). Oligonucleotides for library preparation are shown in the first part of the oligonucleotide table. Reaction conditions for the first PCR were as follows:

Phusion HF buffer (NEB) 5X	5 µl
dNTPs 10 mM	0.5 µl
pDYSCKO_gRNA_for 10 µM	1.25 µl
pDYSCKO_gRNA_rev 10 µM	1.25 µl
Phusion polymerase	0.5 µl
Template DNA (<1 ng/µl)	5 µl
PCR grade water	11.7 µl

496

497 Thermocycler protocol:

Temperature (°C)	Time (s)	Cycles
98	30	1
98	10	16
58	15	
72	5	
72	5	1

498

499 The resulting product was verified on a 2% agarose gel colored with Midori Green Advance
500 (Nippon Genetics, Japan) and then gel-extracted and purified using the FastGene Gel/PCR
501 Extraction Kit (Nippon Genetics, Japan). The purified products were used as the template for the
502 second PCR reaction, with the following conditions:

Phusion Mastermix-HF (NEB)	10 µl
P5-barcode-X oligo 1.333 µM	3.75 µl
P7-barcode-Y oligo 1.333 µM	3.75 µl
Template DNA (~1 ng/µl)	2.5 µl

503

504 Thermocycler protocol:

Temperature (°C)	Time (s)	Cycles
98	30	1
98	10	15
60	10	
72	60	
72	300	1

PCR products were verified on a 2% agarose gel colored with Midori Green Advance (Nippon Genetics, Japan) and then gel-extracted and purified using the FastGene Gel/PCR Extraction Kit (Nippon Genetics, Japan). Library quality control and quantification were performed using the KAPA Library Quantification Kit for Illumina platforms (Kapa Biosystems, Massachusetts, USA) following the manufacturer's instructions. Libraries were then run on a single lane on HiSeq 2500 (Illumina, California, USA) with paired-end 150 bp in fast mode.

Large-scale screen sequencing data analysis

The custom Python scripts used to analyze the data will be made available on github (https://docker.pkg.github.com/Landrylab), packages and software used are presented in Supplementary table 9. Raw sequencing files have been deposited on the NCBI SRA, accession number PRJNA552472. Briefly, reads were separated into three subsequences for alignment: the P5 barcode, the gRNA, and the P7 barcode. Each of these was aligned using Bowtie (Langmead *et al.* 2009) to an artificial reference genome containing either the barcodes or gRNA sequences flanked by the common amplicon sequences. The gRNA sequences are aligned both with 0 or 1 mismatch allowed, and misalignment position and type were stored. Information on barcode and gRNA alignment for each read was stored and combined to generate a barcode count per library table, a list of mismatches in alignments for each gRNA in each library, as well as mismatch types and counts for the same gRNA across all libraries.

Synthesis error within oligonucleotide libraries is one of the major limits of current large-scale genome editing screening methods. These errors can introduce gRNA sequences that cannot perform mutagenesis because the gRNA sequence does not match a site in the genome. We refer to those gRNAs as SE gRNAs. In our experiment, the stringent selection criteria used to select gRNAs limited the risk of off-target effects even for gRNAs with one mismatch, minimizing the risk that a synthesis error gRNA could lead to editing at another site in the genome. We therefore decided to use highly abundant SE gRNAs as negative controls to obtain a null distribution of abundance variation for gRNAs with no fitness effects. To differentiate synthesis errors from sequencing errors, we used the mismatch type and count table to assess whether a particular mismatched gRNA constitutes a too large fraction of the reads associated with a gRNA to be simply a repeated sequencing error. For each error, we test if:

$$\frac{N_{readsformismatch}}{N_{perfectalignment}} > 0.075$$

and discarded the reads associated with the specific mismatch alignment. This threshold was obtained by iteratively testing different threshold values in an effort to maximize the gain in gRNA counts while minimizing the noise added by incorrect assignments. Read counts per library for abundant ($N_{readsformismatch} > 1{,}000$) SE gRNAs were kept to serve as negative controls when measuring fitness effects, resulting in a set of 1,032 abundant SE gRNAs. gRNAs absent from more than half of the libraries (4446 out of 39,989) were removed from the analysis before gRNA abundance calculations.

Detecting mutations with high fitness effects

Barcode sequencing competition experiments use DNA barcodes to measure the relative abundance of many different subpopulations of cells grown in the same pool (Robinson *et al.* 2014). Since each gRNA is linked to its possible mutagenesis outcomes, we can use relative gRNA abundance to detect mutations with significant fitness effects. To do so, the $\log_2$ of the

relative abundance of a barcode after mutagenesis is compared with its abundance at the end of the screen:

$$\Delta\log_{2_{gRNA}} = \log_2\left(\frac{N_{readsgRNAt_1}}{N_{readst_1}}\right) - \log_2\left(\frac{N_{readsgRNAt_0}}{N_{readst_0}}\right)$$

For each gRNA, the measured fitness effect is the product of the effect of the mutational outcomes on growth and of the mutation rate within the cell subpopulation bearing this particular gRNA. Relative counts will also vary stochastically because of variation in sequencing coverage depending on the time point and replicate. To reduce the impact of these effects, a minimal read count at the end of the galactose induction step was used to filter out low abundance gRNAs. We found a minimal read threshold of n=54 provided a good tradeoff between the number of gRNAs eligible for analysis and inter-replicate correlation.

Using the distribution of $\Delta\log 2$ values, we calculated a z-score for each gRNA in both replicates. We then averaged z-scores between replicates and compared the score distributions between SE and Non-SE gRNAs. This revealed the presence of a left-skewed tail in the z-score distribution of valid gRNAs, which is absent in the SE. Because the number of SE gRNAs is smaller than the one of functional gRNAs by almost two orders of magnitude, a type I error (false positives) empirical threshold based solely on a weighted SE z-score distribution was not practical. To resolve this, we fitted a Gumbell left skewed distribution to the SE gRNAs z-score distribution and used it to approximate the type I error rate as a function of the z-score. We set a significance threshold such as that all gRNAs at z-scores for which the estimated false positive rate is below or equal to 5% are considered GNEs.

Complementation assays

Experiments were performed in heterozygous deletion mutants from the YKO project heterozygous deletion strain set (Dharmacon, Colorado, USA). For each gene, a single colony

streaked from the glycerol stock was used to prepare competent cells using the previously described lithium acetate protocol. To generate mutant alleles of the genes of interest, we performed site-directed mutagenesis on the appropriate MoBY collection plasmid (Ho *et al.* 2009). These centromeric plasmids encode the yeast gene of interest under the control of their native promoters and terminators. Mutagenesis reactions were performed with the following reaction setup:

Kapa HiFi buffer (Kapa biosciences) 5X	5 µl
dNTPs 10µM	0.75 µl
mutation_for 10µM (see table 7)	0.75 µl
mutation_rev 10µM (see table 7)	0.75 µl
Kapa Hot-start polymerase	0.5 µl
Template plasmid DNA (15ng/ul)	0.75 µl
PCR grade water	16.5 µl

Thermocycler protocol:

Temperature (°C)	Time (s)	Cycles
95	300	1
98	20	20
60	15	
72	720	
72	1080	1

After amplification, the mutagenesis product was digested with DpnI for 2 hours at 37°C and 5 ul was transformed in *E. coli* strain BW23474 (F-, Δ*(argF-lac)169*, Δ*uidA4::pir-116*, *recA1*, *rpoS396(Am)*, *endA9(del-ins)::FRT*, *rph-1*, *hsdR514*, *rob-1*, *creC510*, Haldimann *et al.* 1996). Transformants were plated on 2YT+Kan+Chlo and grown at 37°C overnight. Plasmid DNA was

585 then isolated from clones and sent for Sanger sequencing (CHUL sequencing platform, Université
586 Laval, Québec City, Canada) to confirm mutagenesis success.

587 Competent cells of target genes were transformed with the appropriate mutant plasmids as well
588 a the original plasmid bearing the wild-type gene and the empty vector (Zhao *et al.* 2016), and
589 transformants were selected by plating on SC-U (MSG). Multiple independent colonies per
590 transformation were then put on sporulation media until sporulation could be confirmed by
591 microscopy. For tetrad dissection, cells were resuspended in 100ul 20T zymolyase (200mg/ml
592 dilution in water) and incubated for 20 minutes at room temperature. Cells were then centrifuged
593 and resuspended in 50ul 1M sorbitol before being streaked on a level YPD plate. All dissections
594 were performed using a Singer SporePlay microscope (Singer Instruments, UK). Plate pictures
595 were taken after five days incubation at room temperature except for the RAP1 plasmid
596 complementation test for which the picture was taken after three days. Pictures are shown in
597 Supplementary image 1.

598

599 **Strain construction for confirmations in *RAP1***

600 Because the MoBY collection plasmid for RAP1 cannot fully complement the gene deletion
601 (Supplementary image file 1), we instead performed confirmations by engineering mutations a
602 diploid strain to create heterozygous mutants. *RAP1* was first tagged with a modified version of
603 fragment DHFR F[1,2] (the first half) of the mDHFR enzyme (Tarassov *et al.* 2008). The
604 mDHFR[1,2]-FLAG cassette was amplified using gene-specific primers and previously described
605 reaction parameters (Tarassov *et al.* 2008). Cells were transformed with the cassette using the
606 previously described transformation protocol and were plated on YPD+Nourseothricine (YPD+Nat
607 in Media table). Positive clones were identified by colony PCR and successful fragment fusion
608 was confirmed by Sanger sequencing (CHUL sequencing platform). We then mated the confirmed

609 clones with strain Y8205 (*Matα can1::STE2pr-his5 lyp1::STE3prLEU2 ∆ura3 ∆his3 ∆leu2*, Kindly
610 gifted by Charlie Boone) by inoculating a 4ml YPD culture with overnight starter cultures of both
611 strains and letting the culture grow overnight. Cells were then streaked on YPD+Nat and diploid
612 cells were identified by colony PCR using mating type diagnosis primers (Huxley *et al.* 1990).

613 To create heterozygous deletion mutants of the target gene, we amplified a modified version of
614 the *URA3* cassettes that could then be targeted with the CRISPR-Cas9 system to integrate our
615 mutations of interest using homologous recombination at the target locus. The oligonucleotides
616 we used differ from those commonly used in that they amplify the cassette without the two LoxP
617 sites present at both ends. We found it necessary to remove those sites as one common
618 mutational outcome after introducing a double-stranded break in the *URA3* cassette was inter-
619 LoxP site recombination without the integration of donor DNA at the target locus. These modified
620 cassettes recombine with DNA upstream the target gene on one end and the mDHFR F[1,2]
621 fusion on the other, ensuring that the heterozygous deletion is always performed at the locus that
622 is already tagged. Cassettes were transformed using the standard lithium acetate method, and
623 cells were plated on SC-U (MSG) selective media. Heterozygous deletion mutants were then
624 confirmed by colony PCR.

625 **CRISPR-Cas9 mediated Knock-in of targeted mutations**

626 Mutant alleles of target genes were amplified in two fragments using template DNA from the
627 haploid tagged strain (See Figure S8). The two fragments bearing mutations are then fused
628 together by a second PCR round to form the final donor DNA. This DNA was then co-transformed
629 with a plasmid bearing Cas9 and a gRNA targeting the URA3 cassette for HDR mediated editing
630 using a standard protocol (Ryan *et al.* 2016). Clones were then screened by PCR to verify donor
631 DNA and mutation integration at the target locus. The targeted region of *RAP1* was then Sanger
632 sequenced (CHUL sequencing platform, Univesité Laval, Québec City, Canada) to confirm the
633 presence of the mutation of interest. Heterozygous mutants were sporulated on solid media until

sporulation could be confirmed by microscopy using the same protocol previously described. The plates were then replica plated on YPD+Nat media, and the pictures were taken after five days at room temperature Supplementary image 2.

Evolutionary rate measurements and protein variant abundance

Evolutionary rates were calculated using the Rate4site software (Mayrose *et al.* 2004) using multiple sequence alignments and phylogenies from PhylomeDB V4 (Huerta-Cepas *et al.* 2014) as input and using the raw calculated rates as output. Variant data was compiled using data from the 1002 Yeast Genome Project (http://1002genomes.u-strasbg.fr/files/allReferenceGenesWithSNPsAndIndelsInferred.tar.gz). Strain-specific protein coding sequence were aligned to the S288c sequence using Fastx36 (Pearson *et al.* 1997) with the following parameters: `fastx36 -p -s -VT10 -T 6 -m 10 -n -3 querymultifasta.fasta ref_orf.db 12\> fasta_out` . Alignments were then parsed with a custom Python script to identify variants. Variant abundance was measured as the number of strains in the dataset in which a specific variant was found. If the coding sequence contained ambiguous nucleotides (ex.: R or Y), separate coding sequences were generated for each possibility and each possible variant was considered as a separate occurrence.

Analysis of the properties of stop codon generating gRNAs

To analyse the sequence and target properties of gRNA inducing the creation of stop codons, data from multiple sources was compiled. For each target gene, length and chromosomal strand was obtained from the Saccharomyces Genome Database using the Yeastmine query interface (Cherry *et al.* 2012). Distance to centromere was obtained by calculating the minimal distance between the start of the gene and one extremity of the centromere coordinates. RNA:DNA duplex melting temperature of gRNA sequence with target genomic DNA was calculated using the MeltingTemp module from Biopython (Cock *et al.* 2009), which uses values taken from Sugimoto

et al (Sugimoto *et al.* 1995). Correlation between gRNA/DNA duplex melting temperatures was assessed using Spearman's rank correlation.

Variant effect prediction resources analysis and GO enrichment

All prediction data except the Envision scores were extracted from the aggregated data of the Mutfunc database (Wagih *et al.* 2018). Precomputed values were downloaded directly from the FTP server (http://ftp.ebi.ac.uk/pub/databases/mutfunc/mutfunc_v1/yeast/). This database includes precomputed SIFT scores for 5498 yeast proteins, as well as predicted variant ddG values based on protein structure (n=1057), homology models (n=1703) and protein-protein interaction interfaces (n=1109). Mutations with $\Delta\Delta G>1$ considered destabilizing.

Precomputed values from Envision (Gray *et al.* 2018) were downloaded directly from the database website (https://envision.gs.washington.edu/shiny/envision_new/, file yeast_predicted_2017-03-12.csv). This file contained 34857830 mutation effect predictions spread across 4011 genes. The distribution of Envision scores for the genes targeted in the experiment that are included in the database are shown in (Figure S6).

Gene enrichments were performed using the PANTHER gene list analysis tool (Mi *et al.* 2019). The list of genes for which 2 or more GNEs were detected was tested for enrichment against all genes targeted by the library using Fisher's exact test and False Discovery Rate calculations. The Gene Ontology datasets used were: GO molecular function complete, GO biological process complete, and GO cellular component complete.

Supplementary dataset 1 contains all gRNAs, their z-scores values as well all the information and annotations used in data analysis.

Acknowledgments

This work was supported by the Canadian Institutes of Health Research Foundation grant 387697 to CRL., as well as project grants 364920, 384483, a Frederick Banting and Charles Best graduate scholarship and a Vanier graduate scholarship to P.C.D, by Université Laval via an André Darveau Fellowship to P.C.D., the Fonds Québécois de Recherche en Santé via a Master's training award to P.C.D. and the Japan Society for the Promotion of Science grant numbers S15734 and S17161 to C.R.L. and N.Y. The authors thank Mathieu Hénault, Johan Hallin, and Dan Yamamoto Evans for comments on the manuscript, as well as Maria Isabel Acosta Lopez for assistance during the strain construction process.

Author contributions

PCD, AKD, NY and CRL designed research. PCD and AKD performed experiments. PCD and MS generated NGS sequencing data. All data analysis was performed by PCD with input from CRL. PCD and CRL wrote the manuscript with input from all authors.

Conflict of interest

None to declare

References

Albuquerque C. P., M. B. Smolka, S. H. Payne, V. Bafna, J. Eng, *et al.*, 2008 A multidimensional chromatography technology for in-depth phosphoproteome analysis. Mol. Cell. Proteomics 7: 1389–96. https://doi.org/10.1074/mcp.M700468-MCP200

Bao Z., M. HamediRad, P. Xue, H. Xiao, I. Tasan, *et al.*, 2018 Genome-scale engineering of Saccharomyces cerevisiae with single-nucleotide precision. Nat. Biotechnol. https://doi.org/10.1038/nbt.4132

C. elegans Deletion Mutant Consortium T. C. elegans D. M., 2012 Large-Scale Screening for Targeted Knockouts in the Caenorhabditis elegans Genome. G3; Genes|Genomes|Genetics 2: 1415–1425. https://doi.org/10.1534/g3.112.003830

Casadaban M. J., and S. N. Cohen, 1980 Analysis of gene control signals by DNA fusion and cloning in Escherichia coli. J. Mol. Biol. 138: 179–207. https://doi.org/10.1016/0022-

2836(80)90283-1

Cherry J. M., E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, *et al.*, 2012 Saccharomyces Genome Database: The genomics resource of budding yeast. Nucleic Acids Res. https://doi.org/10.1093/nar/gkr1029

Cock P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, *et al.*, 2009 Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. https://doi.org/10.1093/bioinformatics/btp163

Copley R. R., and G. J. Barton, 1994 A Structural Analysis of Phosphate and Sulphate Binding Sites in Proteins. J. Mol. Biol. 242: 321–329. https://doi.org/10.1006/jmbi.1994.1583

Dandage R., P. C. Després, N. Yachie, and C. R. Landry, 2019 beditor: A Computational Workflow for Designing Libraries of Guide RNAs for CRISPR-Mediated Base Editing. Genetics 212: 377–385. https://doi.org/10.1534/genetics.119.302089

DePristo M. A., D. M. Weinreich, and D. L. Hartl, 2005 Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Genet. 6: 678–687. https://doi.org/10.1038/nrg1672

Després P. C., A. K. Dubé, L. Nielly-Thibault, N. Yachie, and C. R. Landry, 2018 Double Selection Enhances the Efficiency of Target-AID and Cas9-Based Genome Editing in Yeast. G3 (Bethesda). g3.200461.2018. https://doi.org/10.1534/g3.118.200461

Dicarlo J. E., J. E. Norville, P. Mali, X. Rios, J. Aach, *et al.*, 2013 Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res. 41: 4336–4343. https://doi.org/10.1093/nar/gkt135

Doench J. G., E. Hartenian, D. B. Graham, Z. Tothova, M. Hegde, *et al.*, 2014 Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat. Biotechnol. 32: 1262–1267. https://doi.org/10.1038/nbt.3026

Doench J. G., N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg, *et al.*, 2016 Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. 34: 184–191. https://doi.org/10.1038/nbt.3437

Eriani G., M. Delarue, O. Poch, J. Gangloff, and D. Moras, 1990 Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. Nature 347: 203–206. https://doi.org/10.1038/347203a0

Fowler D. M., and S. Fields, 2014 Deep mutational scanning: a new style of protein science. Nat. Methods 11: 801–7. https://doi.org/10.1038/nmeth.3027

Gaudelli N. M., A. C. Komor, H. A. Rees, M. S. Packer, A. H. Badran, *et al.*, 2017 Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature 551: 464–471. https://doi.org/10.1038/nature24644

Giaever G., D. D. Shoemaker, T. W. Jones, H. Liang, E. A. Winzeler, *et al.*, 1999 Genomic profiling of drug sensitivities via induced haploinsufficiency. Nat. Genet. 21: 278–83. https://doi.org/10.1038/6791

Giaever G., A. M. Chu, L. Ni, C. Connelly, L. Riles, *et al.*, 2002 Functional profiling of the Saccharomyces cerevisiae genome. Nature 418: 387–391. https://doi.org/10.1038/nature00935

749 Gietz R. D., and R. H. Schiestl, 2007 High-efficiency yeast transformation using the LiAc/SS
750 carrier DNA/PEG method. Nat. Protoc. 2: 31–34. https://doi.org/10.1038/nprot.2007.13

751 Graham I. R., R. A. Haw, K. G. Spink, K. A. Halden, and A. Chambers, 1999 In vivo analysis of
752 functional regions within yeast Rap1p. Mol. Cell. Biol. 19: 7481–90.
753 https://doi.org/10.1128/mcb.19.11.7481

754 Grant T. D., J. R. Luft, J. R. Wolfley, M. E. Snell, H. Tsuruta, *et al.*, 2013 The structure of yeast
755 glutaminyl-tRNA synthetase and modeling of its interaction with tRNA. J. Mol. Biol. 425:
756 2480–2493. https://doi.org/10.1016/j.jmb.2013.03.043

757 Grate L., and M. Ares, 2002 Searching yeast intron data at Ares lab web site. Methods
758 Enzymol. https://doi.org/10.1016/S0076-6879(02)50975-7

759 Gray V. E., R. J. Hause, J. Luebeck, J. Shendure, and D. M. Fowler, 2018 Quantitative
760 Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. Cell Syst.
761 https://doi.org/10.1016/j.cels.2017.11.003

762 Haldimann A., M. K. Prahalad, S. L. Fisher, S. K. Kim, C. T. Walsh, *et al.*, 1996 Altered
763 recognition mutants of the response regulator PhoB: a new genetic strategy for studying
764 protein-protein interactions. Proc. Natl. Acad. Sci. U. S. A. 93: 14361–6.
765 https://doi.org/10.1073/pnas.93.25.14361

766 Hess G. T., L. Frésard, K. Han, C. H. Lee, A. Li, *et al.*, 2016 Directed evolution using dCas9-
767 targeted somatic hypermutation in mammalian cells. Nat. Methods.
768 https://doi.org/10.1038/nmeth.4038

769 Ho C. H., L. Magtanong, S. L. Barker, D. Gresham, S. Nishimura, *et al.*, 2009 A molecular
770 barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. Nat.
771 Biotechnol. 27: 369–377. https://doi.org/10.1038/nbt.1534

772 Holt L. J., B. B. Tuch, J. Villén, A. D. Johnson, S. P. Gygi, *et al.*, 2009 Global analysis of Cdk1
773 substrate phosphorylation sites provides insights into evolution. Science 325: 1682–6.
774 https://doi.org/10.1126/science.1172867

775 Huerta-Cepas J., S. Capella-Gutiérrez, L. P. Pryszcz, M. Marcet-Houben, and T. Gabaldón,
776 2014 PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome.
777 Nucleic Acids Res. https://doi.org/10.1093/nar/gkt1177

778 Huxley C., E. D. Green, and I. Dunham, 1990 Rapid assessment of S. cerevisiae mating type by
779 PCR. Trends Genet. 6: 236.

780 Jin S., Y. Zong, Q. Gao, Z. Zhu, Y. Wang, *et al.*, 2019 Cytosine, but not adenine, base editors
781 induce genome-wide off-target mutations in rice. Science eaaw7166.
782 https://doi.org/10.1126/science.aaw7166

783 Kim D., K. Lim, S. T. Kim, S. H. Yoon, K. Kim, *et al.*, 2017 Genome-wide target specificities of
784 CRISPR RNA-guided programmable deaminases. Nat. Biotechnol.
785 https://doi.org/10.1038/nbt.3852

786 Konig P., R. Giraldo, L. Chapman, and D. Rhodes, 1996 The crystal structure of the DNA-
787 binding domain of yeast RAP1 in complex with telomeric DNA. Cell 85: 125–36.
788 https://doi.org/10.1016/S0092-8674(00)81088-0

789 Landry C. R., E. D. Levy, and S. W. Michnick, 2009 Weak functional constraints on
790 phosphoproteomes. Trends Genet. 25: 193–7. https://doi.org/10.1016/j.tig.2009.03.003

Langmead B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10: R25. https://doi.org/10.1186/gb-2009-10-3-r25

Ma Y., J. Zhang, W. Yin, Z. Zhang, Y. Song, *et al.*, 2016 Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. Nat. Methods 13: 1029–1035. https://doi.org/10.1038/nmeth.4027

Mayrose I., D. Graur, N. Ben-Tal, and T. Pupko, 2004 Comparison of site-specific rate-inference methods for protein sequences: Empirical Bayesian methods are superior. Mol. Biol. Evol. https://doi.org/10.1093/molbev/msh194

Mi H., A. Muruganujan, X. Huang, D. Ebert, C. Mills, *et al.*, 2019 Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat. Protoc. 14: 703–721. https://doi.org/10.1038/s41596-019-0128-8

Michel A. H., R. Hatakeyama, P. Kimmig, M. Arter, M. Peter, *et al.*, 2017 Functional mapping of yeast genomes by saturated transposition. Elife 6. https://doi.org/10.7554/eLife.23570

Ng P. C., and S. Henikoff, 2003 SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res.

Nishida K., T. Arazoe, N. Yachie, S. Banno, M. Kakimoto, *et al.*, 2016 Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. Science (80-.). 353: 553–563. https://doi.org/10.1126/science.aaf8729

Nishimasu H., X. Shi, S. Ishiguro, L. Gao, S. Hirano, *et al.*, 2018 Engineered CRISPR-Cas9 nuclease with expanded targeting space. Science eaas9129. https://doi.org/10.1126/science.aas9129

Pearson W. R., T. Wood, Z. Zhang, and W. Miller, 1997 Comparison of DNA Sequences with Protein Sequences. Genomics 46: 24–36. https://doi.org/10.1006/geno.1997.4995

Qi L. S., M. H. Larson, L. A. Gilbert, J. A. Doudna, J. S. Weissman, *et al.*, 2013 Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell 152: 1173–83. https://doi.org/10.1016/j.cell.2013.02.022

Qian W., D. Ma, C. Xiao, Z. Wang, and J. Zhang, 2012 The Genomic Landscape and Evolutionary Resolution of Antagonistic Pleiotropy in Yeast. Cell Rep. 2: 1399–1410. https://doi.org/10.1016/j.celrep.2012.09.017

Rees H. A., and D. R. Liu, 2018 Base editing: precision chemistry on the genome and transcriptome of living cells. Nat. Rev. Genet. 19: 770–788. https://doi.org/10.1038/s41576-018-0059-1

Reis A. M. C., W. K. Mills, I. Ramachandran, E. C. Friedberg, D. Thompson, *et al.*, 2012 Targeted detection of in vivo endogenous DNA base damage reveals preferential base excision repair in the transcribed strand. Nucleic Acids Res. 40: 206–219. https://doi.org/10.1093/nar/gkr704

Roy K. R., J. D. Smith, S. C. Vonesch, G. Lin, C. S. Tu, *et al.*, 2018 Multiplexed precision genome editing with trackable genomic barcodes in yeast. Nat. Biotechnol. https://doi.org/10.1038/nbt.4137

Ryan O. W., S. Poddar, and J. H. D. Cate, 2016 Crispr–cas9 genome engineering in Saccharomyces cerevisiae cells. Cold Spring Harb. Protoc. 2016: 525–533.

https://doi.org/10.1101/pdb.prot086827

Sander J. D., and J. K. Joung, 2014 CRISPR-Cas systems for editing, regulating and targeting genomes. Nat. Biotechnol. 32: 347–55. https://doi.org/10.1038/nbt.2842

Sanjana N. E., O. Shalem, and F. Zhang, 2014 Improved vectors and genome-wide libraries for CRISPR screening. Nat. Methods 11: 783–784. https://doi.org/10.1038/nmeth.3047

Schmitt E., M. Panvert, S. Blanquet, and Y. Mechulam, 1995 Transition state stabilization by the "high" motif of class I aminoacyl-tRNA synthetases: The case of Escherichia coli methionyl-tRNA synthetase. Nucleic Acids Res. https://doi.org/10.1093/nar/23.23.4793

Schymkowitz J., J. Borg, F. Stricher, R. Nys, F. Rousseau, *et al.*, 2005 The FoldX web server: an online force field. Nucleic Acids Res. 33: W382–W388. https://doi.org/10.1093/nar/gki387

Shalem O., N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, *et al.*, 2014 Genome-scale CRISPR-Cas9 knockout screening in human cells. Science (80-.). 343: 84–87. https://doi.org/10.1126/science.1247005

Sharon E., S. A. A. Chen, N. M. Khosla, J. D. Smith, J. K. Pritchard, *et al.*, 2018 Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. Cell. https://doi.org/10.1016/j.cell.2018.08.057

Smith J. E., J. R. Alvarez-Dominguez, N. Kline, N. J. Huynh, S. Geisler, *et al.*, 2014 Translation of Small Open Reading Frames within Unannotated RNA Transcripts in Saccharomyces cerevisiae. Cell Rep. 7: 1858–1866. https://doi.org/10.1016/j.celrep.2014.05.023

Smith J. D., S. Suresh, U. Schlecht, M. Wu, O. Wagih, *et al.*, 2016 Quantitative CRISPR interference screens in yeast identify chemical-genetic interactions and new rules for guide RNA design. Genome Biol. 17: 45. https://doi.org/10.1186/s13059-016-0900-9

Sneath P. H., 1966 Relations between chemical structure and biological activity in peptides. J. Theor. Biol. 12: 157–95.

Sugimoto N., S. Nakano, M. Katoh, A. Matsumura, H. Nakamuta, *et al.*, 1995 Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. Biochemistry 34: 11211–6.

Tarassov K., V. Messier, C. R. Landry, S. Radinovic, M. M. Serna Molina, *et al.*, 2008 An in vivo map of the yeast protein interactome. Science 320: 1465–70. https://doi.org/10.1126/science.1153878

Wagih O., M. Galardini, B. P. Busby, D. Memon, A. Typas, *et al.*, 2018 A resource of variant effect predictions of single nucleotide variants in model organisms. Mol. Syst. Biol. 14: e8430. https://doi.org/10.15252/MSB.20188430

Winzeler E. A., D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, *et al.*, 1999 Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science (80-.). 285: 901–906. https://doi.org/10.1126/science.285.5429.901

Wong N., W. Liu, and X. Wang, 2015 WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. Genome Biol. 16: 218. https://doi.org/10.1186/S13059-015-0784-0

Wu A. C. K., H. Patel, M. Chia, F. Moretto, D. Frith, *et al.*, 2018 Repression of Divergent Noncoding Transcription by a Sequence-Specific Transcription Factor. Mol. Cell 72: 942-

954.e7. https://doi.org/10.1016/J.MOLCEL.2018.10.018

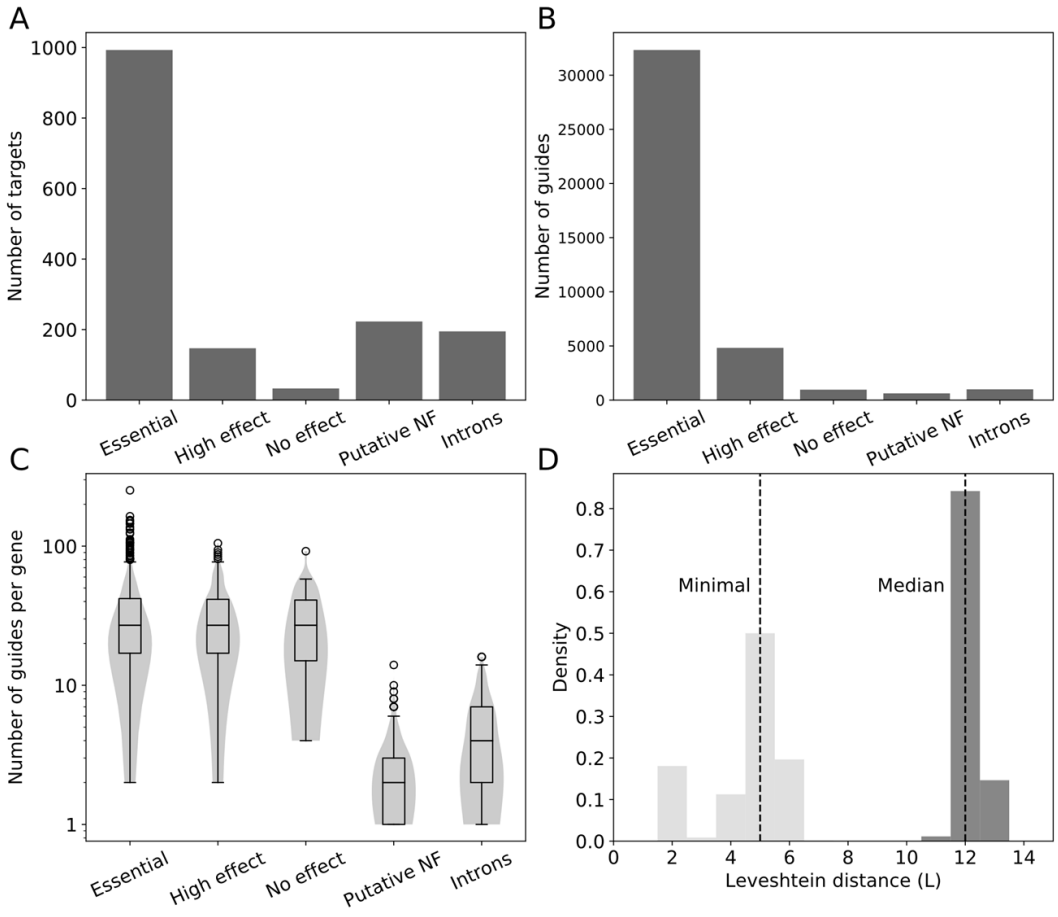
Yachie N., E. Petsalaki, J. C. Mellor, J. Weile, Y. Jacob, *et al.*, 2016 Pooled-matrix protein interaction screens using Barcode Fusion Genetics. Mol. Syst. Biol. 12: 863.

Zhao L., Q. Yang, J. Zheng, X. Zhu, X. Hao, *et al.*, 2016 A genome-wide imaging-based screening to identify genes involved in synphilin-1 inclusion formation in Saccharomyces cerevisiae. Sci. Rep. 6: 30134. https://doi.org/10.1038/srep30134

Zuo E., Y. Sun, W. Wei, T. Yuan, W. Ying, *et al.*, 2019 Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. Science (80-.). eaav9973. https://doi.org/10.1126/SCIENCE.AAV9973

884 SUPPLEMENTARY MATERIAL: Supplementary Figures 1-11

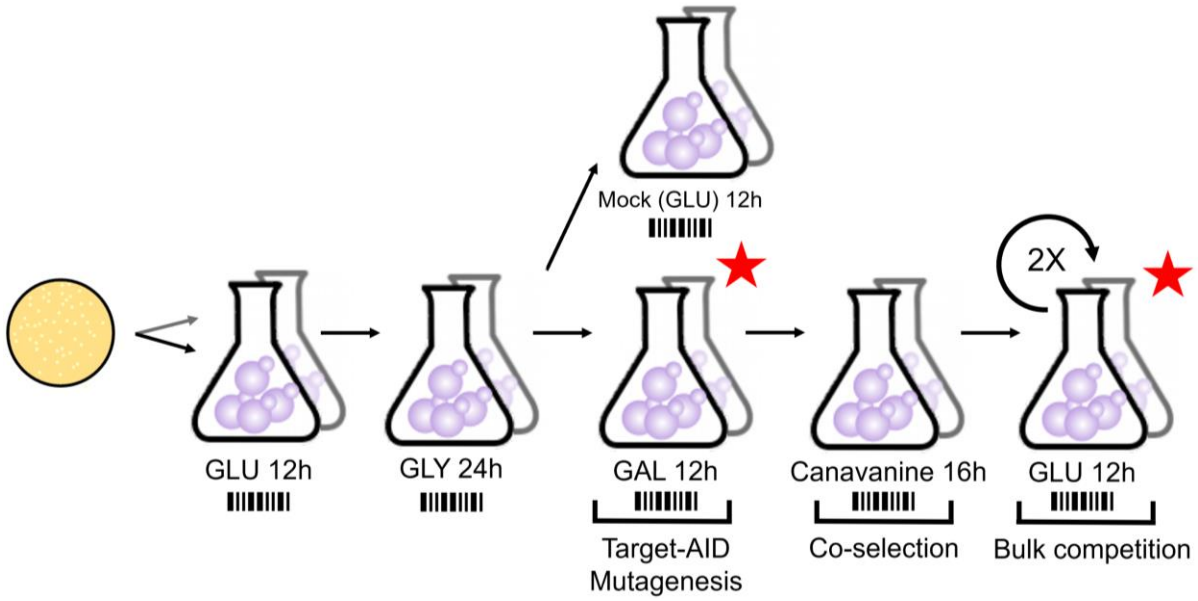
885



886

887 **Figure S1: A gRNA library for the systematic mutagenesis of yeast essential genes and**
888 **other targets of interest. A)** Number of genes targeted by the gRNA library for the different target
889 classes. **B)** Total number of gRNAs targeting genes in the different target classes. **C)** Distribution
890 of number gRNAs for each gene targeted in the different classes. **D)** Distribution of minimal (light
891 grey) and median (dark grey) pairwise sequence distance between all gRNA sequences in the
892 library.

893



894

895 **Figure S2: Experimental workflow for Target-AID mutagenesis and co-selection.** The
896 mutagenesis method closely follows the base editing protocol previously described (Després *et*
897 *al.* 2018). After a pooled transformation step, cells were scraped and splitted into two replicates
898 for pre-cultures. After each step of the protocol, plasmid DNA was extracted from a cell sample
899 and used to amplify and sequence the gRNA pool. The red stars indicate time points used for
900 fitness effects analysis: read counts after galactose induction were used as T0 and were
901 compared with read counts after two rounds of competition. The mock induction steps mimics the
902 induction conditions but galactose in the media is replaced by glucose. This prevents the editing
903 enzyme from being expressed because glucose represses the GAL pathway. After canavanine
904 co-selection, cells go through two competition rounds in synthetic media were selective pressure
905 for the Target-AID bearing plasmid is lost. The entire experiment was completed within less than
906 25 generations after galactose induction, limiting the impact of compensatory and spontaneous
907 mutations.

908

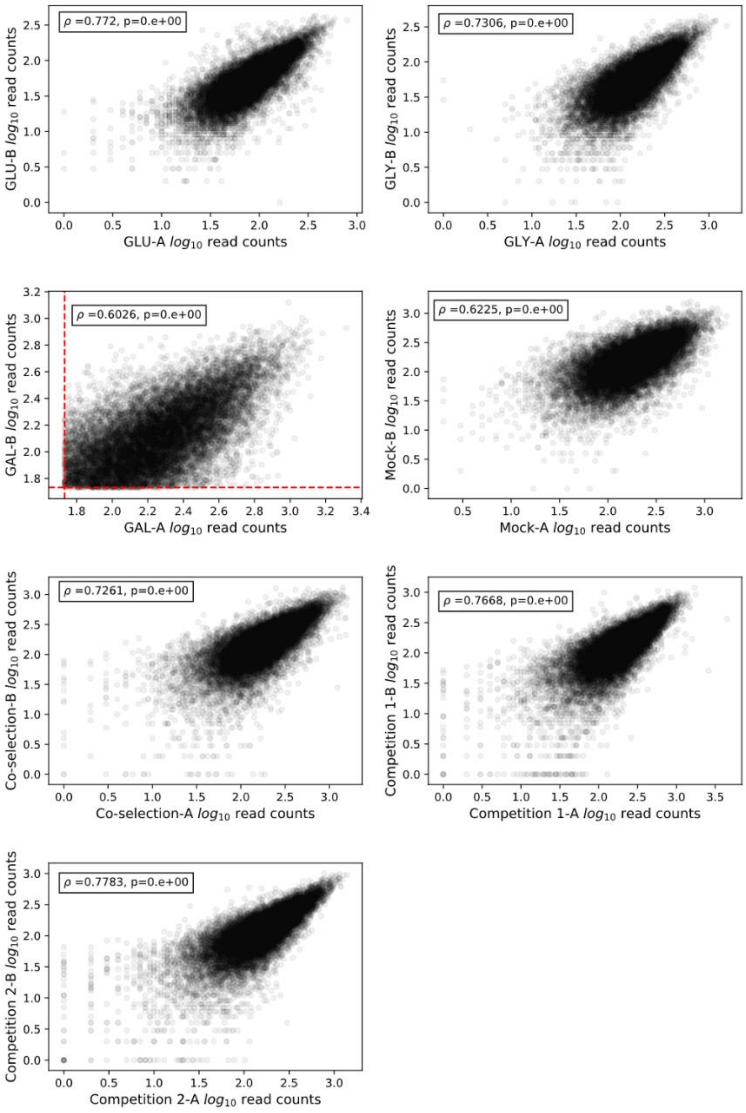
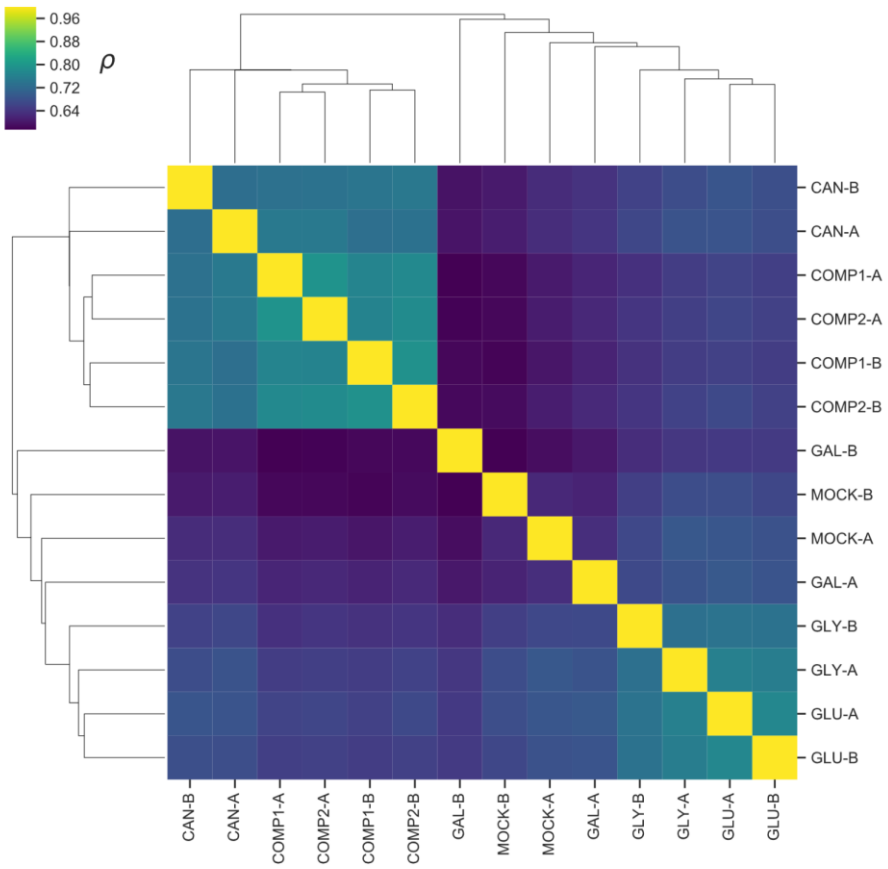


Figure S3: Read abundance rank order is strongly correlated between replicates. For each time point, Spearman rank correlation of gRNA $\log_{10}$ read abundance after basic filtering is shown. The minimal read count after galactose induction, which served as the principal filtering criteria, is shown on the galactose subpanel.



914

915 **Figure S4: Barcode abundance correlation clusters different experimental steps of the**
916 **screen.** Pairwise Spearman rank correlation of barcode counts was used to cluster the libraries
917 obtained at the different time points described in Figure S2. The lower level of correlation between
918 the galactose induction and mock induction timepoints compared to other associated steps could
919 reflect higher stochasticity in growth caused by cell to cell variation in the metabolic switch from
920 glycerol to sugars as the main carbon source as well as editing in the case of the galactose
921 timepoint.

922

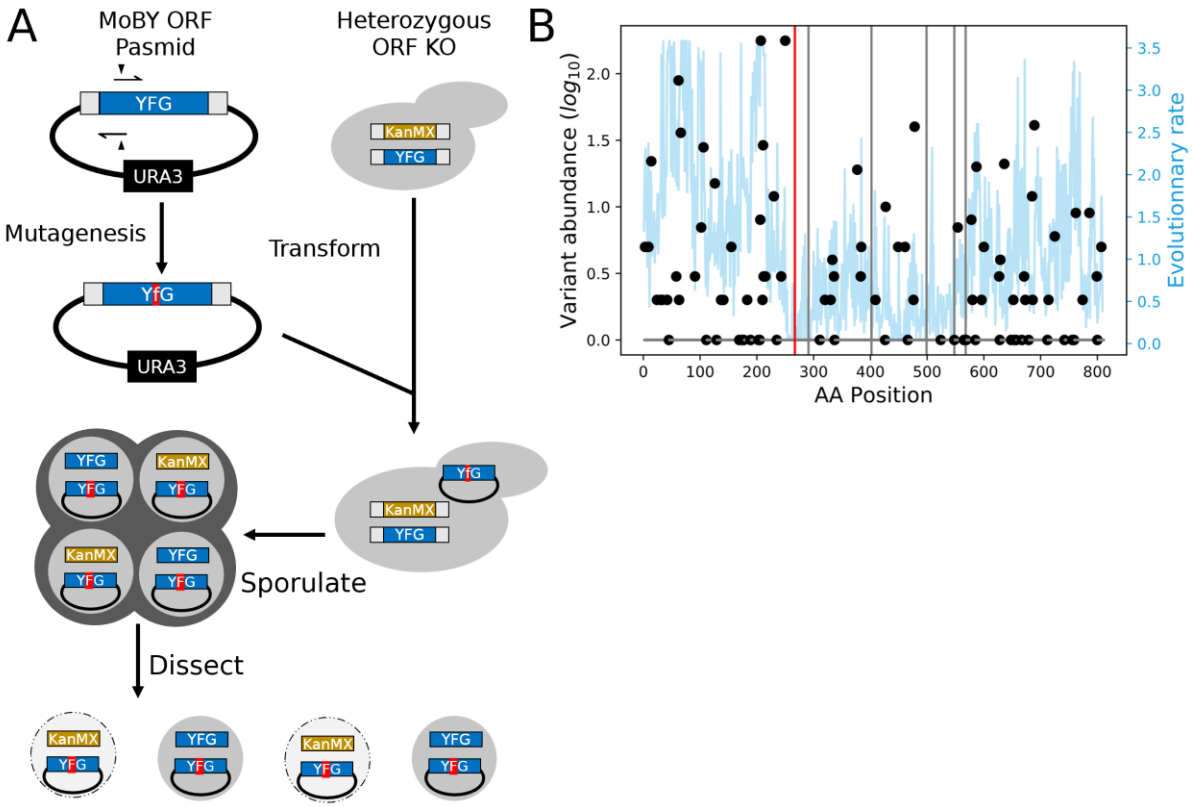


Figure S5 Plasmid-based confirmation workflow by complementation test and evolutionary information on *GLN4*. A) Detailed protocols for the different steps are presented in the methods. First, directed mutagenesis is used to introduce the mutation of interest (shown in red) in the MoBY collection plasmid of the targeted gene (YFG). This vector is then transformed into the heterozygous collection deletion strain (BY4743, *MATa*/α *his3Δ1/his3Δ1 leu2Δ0/leu2Δ0 LYS2/lys2Δ0 met15Δ0/MET15 ura3Δ0/ura3Δ0*) of the gene of interest. The transformants are sporulated and their tetrads are dissected. If the mutated allele carried by the plasmid cannot complement the gene deletion, then only the two progenies bearing the wild-type copies will be viable. **B)** Protein variant frequency among 1000 yeast isolates (black dots) and residue evolutionary rate across species (blue line) for *GLN4*. The target site for the most deleterious GNE is highlighted by a red line and other GNE target sites are shown as grey lines.

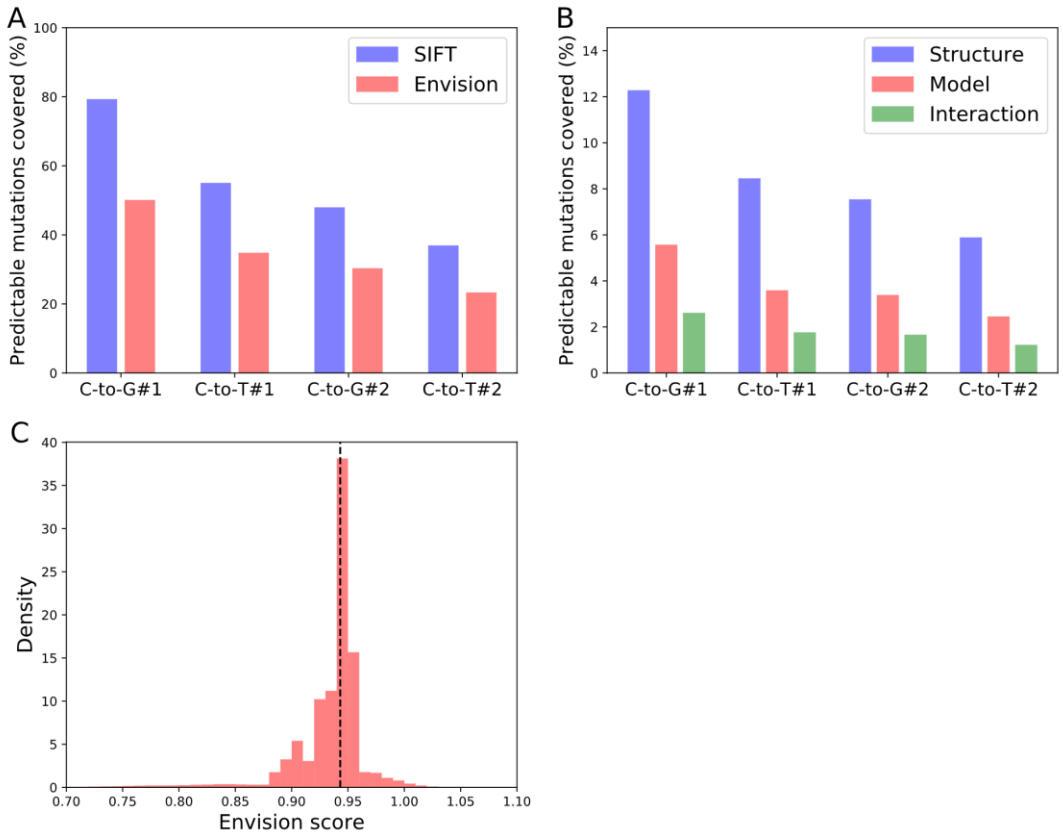


Figure S6 gRNA predicted mutation coverage for Mutfunc and Envision data. Mutfunc integrates both the SIFT prediction scores and FoldX (Schymkowitz *et al.* 2005), ΔΔG predictions for solved protein structures, homology models, and protein-protein interaction interfaces. gRNAs which do not generate missense mutations were included in the calculations. **A)** Coverage for the SIFT and Envision variant effect predictors for the four most probable single mutants created by gRNAs detected in the experiment. **B)** Coverage for ΔΔG predictions for solved protein structures, homology models, and protein-protein interaction interfaces for the four most probable single mutants created by gRNAs detected in the experiment. **C)** Distribution of Envision scores across all sites in the database for all proteins targeted by the set of gRNAs detected in the screen (n=7,556,573). The median score is shown as a dotted black line.

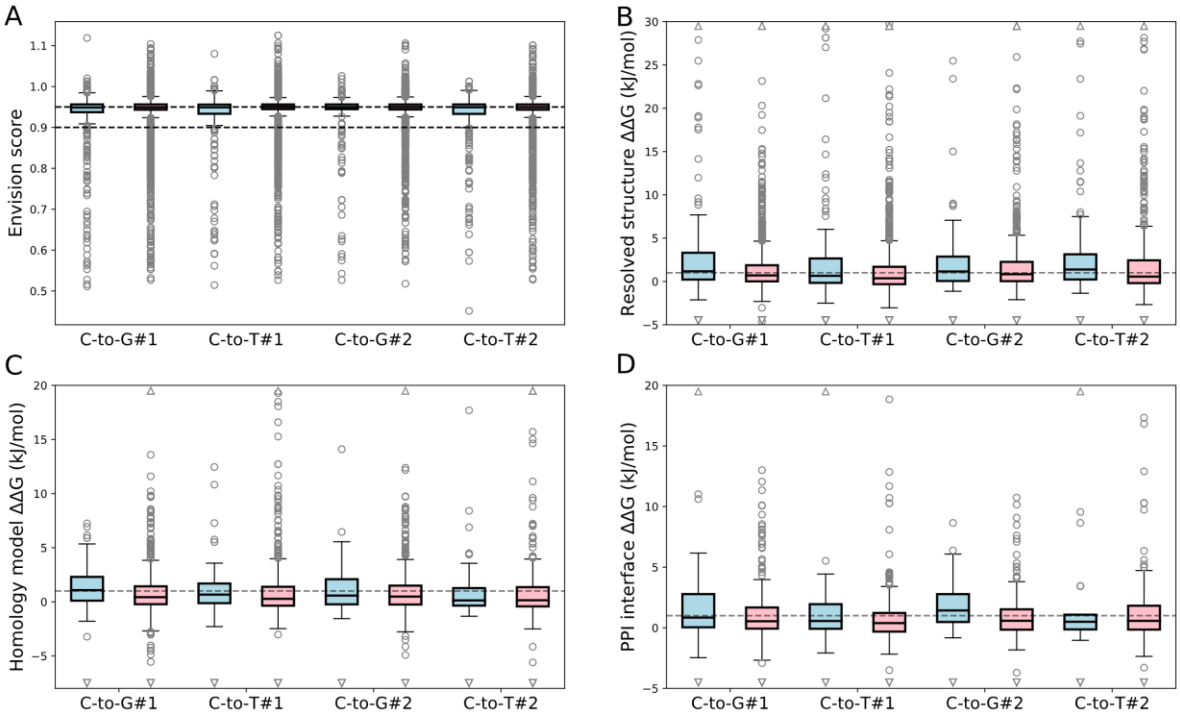


Figure S7 GNE and non-significant gRNA effect prediction distributions. A) Envision score distributions for the four most probable mutations induced by GNEs (blue) and NSGs (red). Welch's t-test p-values for comparisons: 5.00×10^{-6}, 0.002, 0.007, 7.75×10^{-5}. **B)** Predicted folding energy variation (ΔΔG) of GNE and NSG induced protein mutants compared to the wild-type structure based on resolved protein structure. Welch's t-test p-values for comparisons: 0.0001, 0.006, 0.148, 0.007. **C)** Predicted folding energy variation (ΔΔG) of GNE and NSG induced protein mutants compared to the wild-type structure based on homology models of protein structure. Welch's t-test p-values for comparisons: 0.016, 0.441, 0.195, 0.689. **D)** Binding energy variation (ΔΔG) of GNE and NSG induced mutant protein-protein interfaces compared to the wild-type based on a resolved structure on the interface. Welch's t-test p-values for comparisons: 0.285, 0.303, 0.033, 0.95.

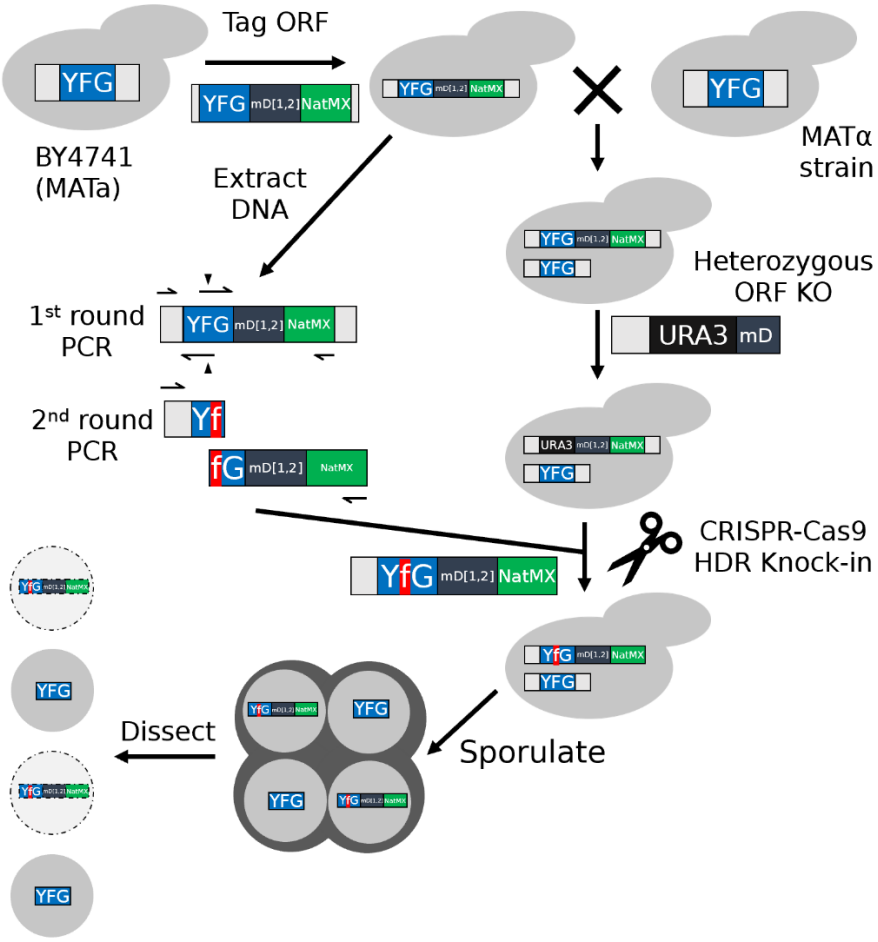
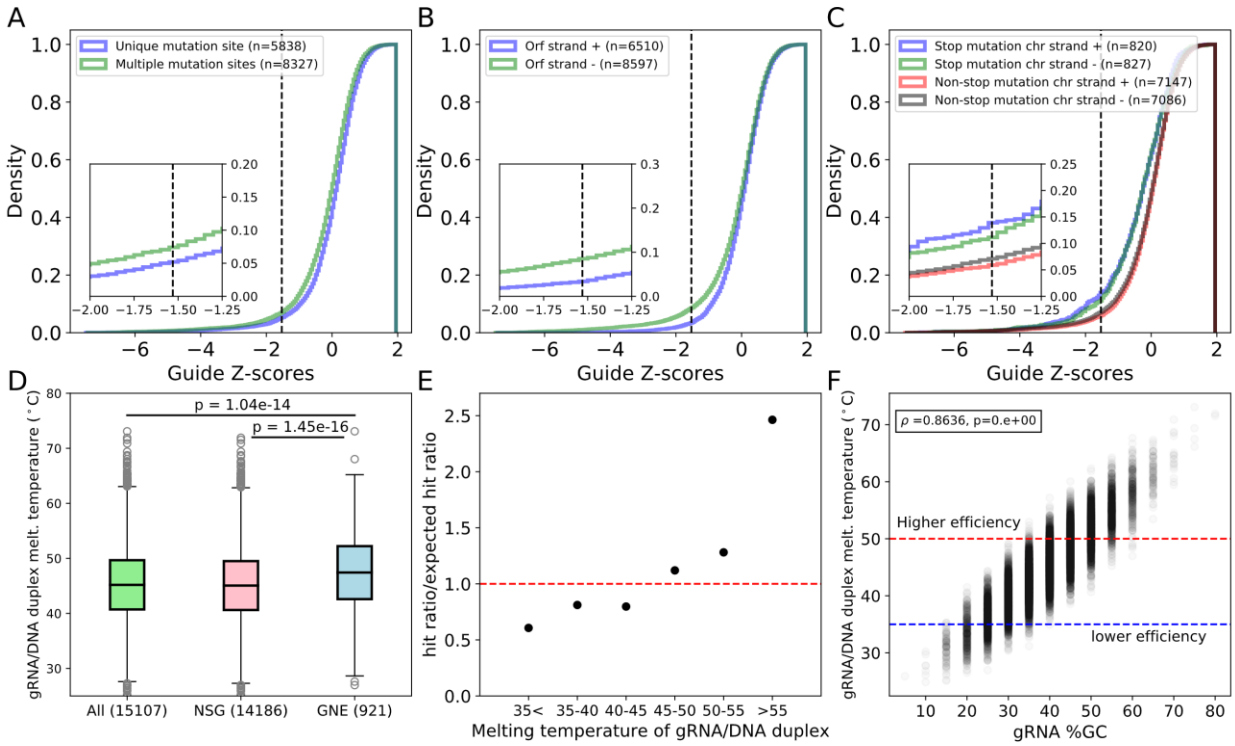


Figure S8 Fitness affecting variant by CRISPR knock in confirmation workflow. Detailed protocols for the different steps are presented in the methods. Starting from the wild-type laboratory strain BY4741, the gene of interest (*YFG*, blue) is first tagged with a modified version of the DHFR F[1,2] cassette (dark gray and green). The tagged strain is then crossed with a MAT*α* strain (Y8205) to create a heterozygous diploid. A *URA3* deletion cassette (black) that recombines with the *YFG* upstream sequence and the start of the mDHFR fragment is then used to generate a heterozygous KO strain. In parallel, genomic DNA is extracted from the tagged haploid strain. This DNA is then used as a template to amplify two fragments of *YFG* bearing the mutation of interest (shown in red) using a set of overhanging primers. The two fragments are then combined by fusion PCR to obtain the donor DNA used in the next step. Using a modified Cas9 vector (Ryan *et al.* 2016) that expresses a gRNA targeting the *URA3* cassette, the mutated allele is introduced at the KO locus to create a heterozygous mutant strain. The diploid cells can then be sporulated, and tetrad dissection allows observation of any phenotype linked with the mutation of interest.



977

978 **Figure S9 Properties influencing stop codon GNEs are generalizable to non-stop codon**
979 **generating GNEs. A)** Cumulative z-score density for gRNAs that do not generate stop codons
980 depending on the number of mutable sites. A higher rate of GNE is observed for gRNAs which
981 can lead to the editing of multiple nucleotides (Two-sample Kolmogorov-Smirnov test, $p=3.91x10^{-29}$).
982 The significance threshold is shown as a black dotted line. **B)** Cumulative z-score density for
983 NSGs on orf target strand. gRNAs targeting the non-coding strand of the ORF have a higher
984 likelihood of being GNEs (Two-sample Kolmogorov-Smirnov test, $p=1.87x10^{-16}$). **C)** gRNA z-score
985 cumulative density for both SGGs and non-SGGs grouped by the chromosomal strand they target.
986 In SGGs, the target strand does not impact z-score distributions (Two-sample Kolmogorov-
987 Smirnov test, p=0.753) and GNE proportions (Fisher's exact test, p=0.149). For non-SGGs, the
988 chromosomal strand has a small influence on z-score distributions (Two-sample Kolmogorov-
989 Smirnov test, p=0.035) and GNE proportions (Fisher's exact test, p=0.002) **D)** Distributions of
990 modeled RNA/DNA duplex melting temperature for all non-SGGs generating gRNAs, the NSG
991 subset, and the GNEs subset. P-values were calculated using the two-sample Kolmogorov-
992 Smirnov test. **E)** Non-SGGs GNE enrichment compared to the expected GNE ratio for different
993 melting temperature ranges. **F)** gRNA/DNA duplex melting temperature as a function of gRNA
994 GC content for all gRNAs for which fitness effects were measured. The higher and lower efficiency
995 thresholds are based on the enrichments shown in panel E and Figure 4F.

996

997

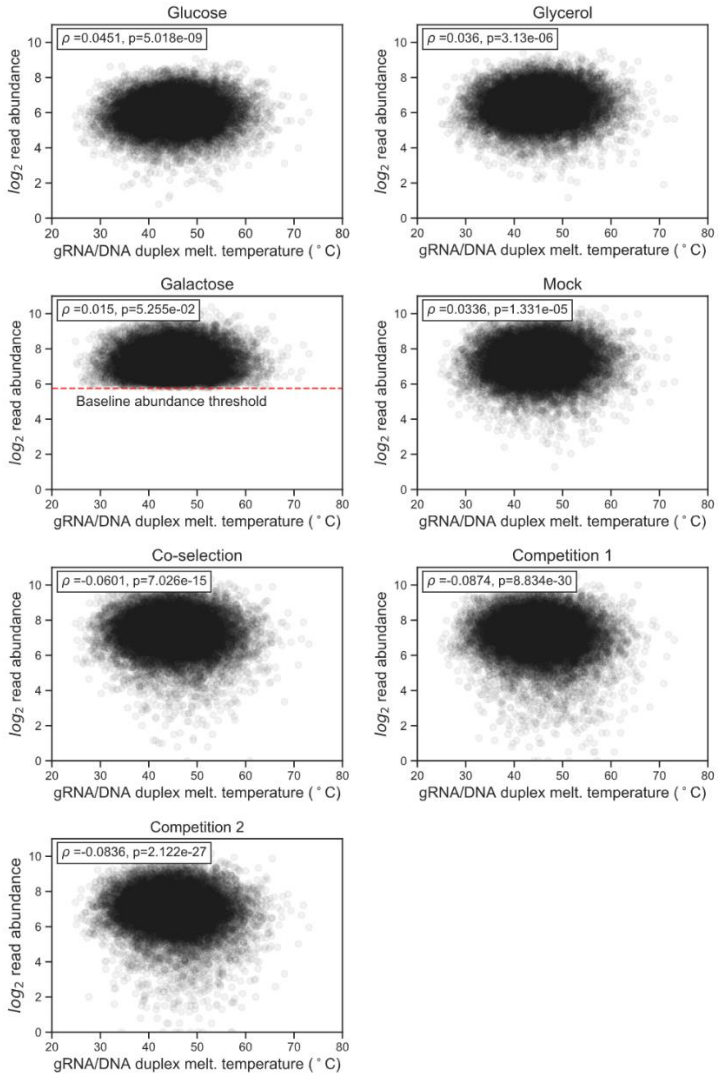


Figure S10 gRNA/DNA duplex melting temperature is not linked to systematic sequencing biases. Spearman rank correlation between replicate averaged read count and predicted gRNA/DNA duplex melting temperature is shown across timepoints. The minimal read count after galactose induction, which served as a filtering criterion, is shown on the galactose subpanels. gRNAs for which no reads were detected in one of the time points were included when computing the correlation but are not shown on the graphs because of log scaling.

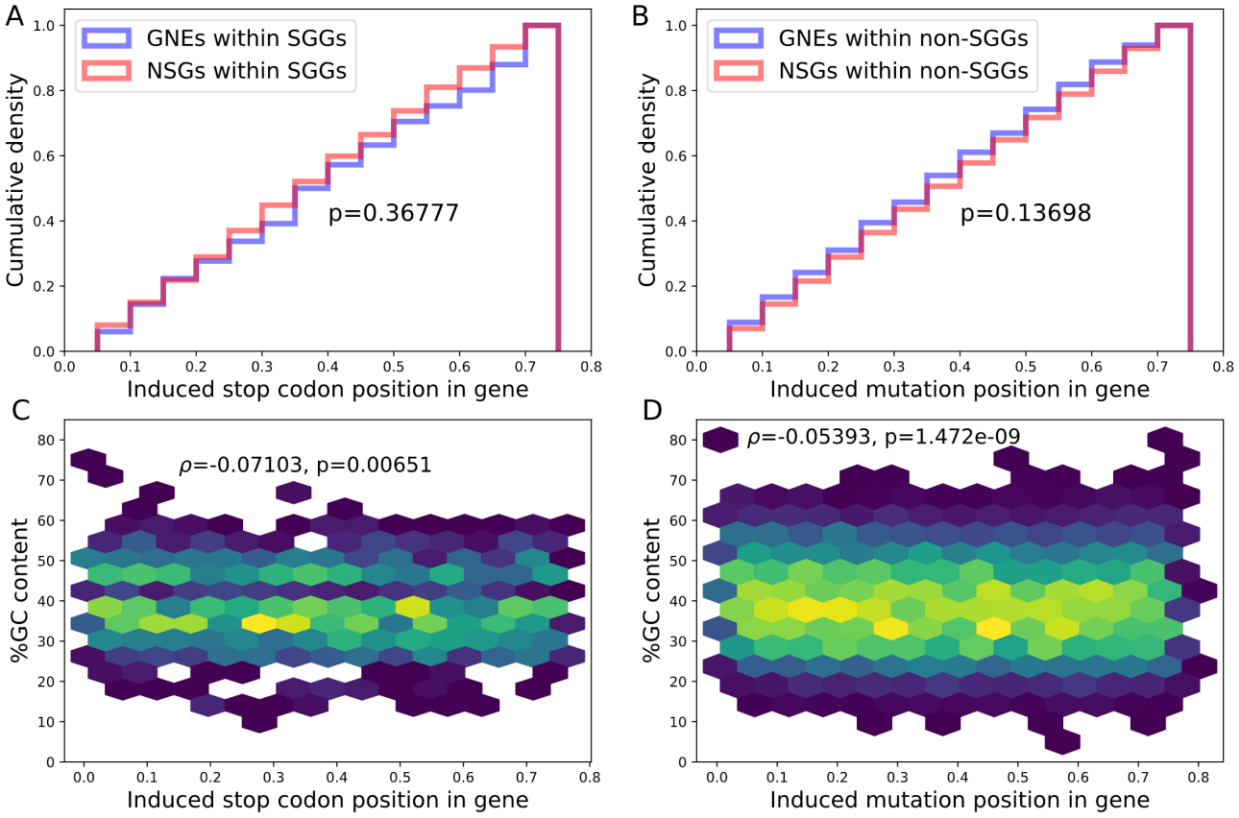


Figure S11 GNE density is independent of target nucleotide position bias. A) In SGGs, GNE and NSG target sites that are evenly distributed across the target genes, and GNEs do not show any bias (Two-sample Kolmogorov-Smirnov). **B)** Non-SGG GNEs do not show any positional bias. **C)** A significant but small negative correlation is observed between gRNA target relative position and GC content of SGGs (Spearman's rank correlation). The very small observed effect coupled with the absence of position bias suggests that relative target position bias does not drive the link between GC content and gRNA efficiency. **D)** Similarly, a small but significant but small negative correlation is also observed between gRNA relative position and GC content for non-SGGs (Spearman's rank correlation).