Perturbing proteomes at single residue resolution 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[1,2]. 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 mutational 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[3–5]. 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[6] or by modifying gene expression levels[7–9] 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[10–14]. 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[15–17]. 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)[18] and CRISPR-X[19], 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[15], 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

Design of a base editing library targeting essential genes

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. Because all essential genes have the same qualitative fitness effects when deleted[20], focusing on these genes allowed us 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 within a locus. We excluded 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[21,22].

To associate each gRNA in the library to specific base editing outcomes, we developed a simple model based on the yeast data included in the original Target-AID manuscript as well as our own work[15,23]. First, we expected that editing would mostly result in genotypes where only one nucleotide is edited in the activity window of the editor. Second, we predicted that the editing outcomes would mainly consist of C to G and C to T mutations and that the abundance of C to A products will be negligible. Finally, we expected that editing frequency ranks would follow the editing activity rankings already known from the initial characterization of Target-AID. Based on these criteria, we filtered out potential target sites where all three high editing rate positions (-19,-18 and -17) or those where both position -18 and -17 are cytosines and kept the remaining sites for inclusion in the gRNA library. The resulting library contained 40 000 gRNAs, of which ~35 000 targeted essential gene coding sequences and ~5000 other target types as shown in Supplementary Figure 1.

Over 75% of target sequences in this set contained only one or two Cs in the extended activity window (positions -20 to -14), and as expected a general enrichment for cytosines in the high activity window (Supplementary Figure 2A-B). Because the goal of our experiment was to link specific mutations to fitness effects, co-editing of multiple nucleotides using an editor which does not channel mutations to a specific outcome has the potential to obscure the genotype responsible for a fitness effect. To take this into account, we placed each gRNA in a co-editing risk category based on the presence and positions of cytosines in the activity window (See methods). Based on this metric, we found that over 80% of gRNAs fell either in the very low or

low risk category (Supplementary Figure 2C). If co-editing occurs, but the other mutated cytosine is part of the same codon as the intended target site, then any resulting fitness effects can still be linked to the perturbation of a specific amino acid. As Target-AID is known to perform processive editing, a high co-editing risk might also be linked to higher overall editing rate[15]. We found the proportion of gRNAs in the library for which this is true to be over 50%: when co-editing risk category is taken into account, the proportion reaches ~90% (Supplementary Figure 2D).

Measurement of mutagenesis rate and outcomes of library gRNAs

While the repair product outcomes of edits for gRNAs can be predicted with varying levels of accuracy for CRISPR-Cas9-based editing[24], no such tools are available yet for base editing applications. As such, the model we used to associate gRNAs in our library to mutational outcomes is only a parsimonious deduction based on the original Target-AID data and our previous work[15,23]. Furthermore, evaluating the activity of gRNAs for base editing remains difficult[25]. The measurement of fitness effects is not associated with a direct simultaneous measurement of mutagenesis rate in our experiment. As such, the absence of fitness effects for a gRNA can both be explained by either non-functional or low editing, or successful editing that resulted in mutations with no detectable fitness effects[23]. As our experiment focuses on the impact of targeted mutations on cell growth, the first group can be seen as false negatives, and the second as true negatives. While we can modulate the gRNA abundance variation threshold to minimize the risk of false positives, additional experimental data on mutagenesis success rates and editing outcomes was required to assess which type of negative results would be dominant in our experiment.

To evaluate the performance of our model and the functionality of the library gRNAs, we performed a base editing time course experiment where mutagenesis rates and outcomes were measured by deep sequencing of the edited genomic loci (Supplementary Figure 3). To gain insights on the mutagenesis outcomes of different editing scenarios, we selected guides with different predicted patterns of cytosine presence in the Target-AID activity window (Figure 1A). We included 9 guides from the library isolated from the library quality control process, as well as three control gRNAs respectively targeting the pseudogene YCL074W, the non-essential gene *VPS17*, and *ADE1*, which can be used as a phenotypic marker. Most gRNAs could efficiently edit their respective targets, with 9 out 12 gRNAs reaching mutation rates of 50% or higher (Figure 1B), consistent with previous results[15,23]. Replicates were highly correlated along different measurements with editing rates at the *CAN1* co-editing site being highly consistent (Supplementary Figure 4A-E). Only the gRNA targeting *SES1* was found to be inactive, and as such was excluded from downstream analysis. The very low editing rate observed for the gRNA targeting *SES1* is an example of unknown factors affecting mutagenesis efficiency that leads to false negatives in large-scale experiments.

In our editing model, we first predict that single mutants would be the main mutagenesis outcome of the base editing process. We found this to be true for 9 gRNAs out of 10 with more than one cytosine in the Target-AID activity window (Figure 1C). Second, our model considers C to A editing to be rare and thus disregards them in favor of the more common C to G and C to T mutations. We observe this bias in the deep sequencing data (Figure 1D), with the median occupancy of both C to G and C to T genotypes in edited alleles being much greater than C to A occupancy (C to T vs C to A: W=0, p=1.73x10^{-6}, C to G vs C to A: W=41, p=8.19x10^{-5}, two-sided wilcoxon signed rank test). Including these mutations as in our model leads to a median coverage of 93% of mutagenesis outcomes. Our sequencing data also showed a greater prevalence of C to T mutations compared to C to G (W=112, p=0.01), but if absolute editing rate is taken

into account this difference disappears (Supplementary Figure 4F). Finally, in cases where multiple editable nucleotides are present in the activity window of the base editor, our model uses the quantitative data of the original Target-AID manuscript to predict qualitatively which position should be edited at the highest frequency. We found that this prediction method of editing rank in the activity window matched with the experimental data in most cases (Figure 1E) which is unlikely to occur by chance ($p \approx 0.0004$ based on $1x10^6$ random rank permutations). Globally, we found that the edited allele pool was mostly composed of the genotypes predicted by our model: for the 8 gRNAs with editing activity that came from the library, the median fraction of edited reads covered by our model was 69% (Figure 1F). In 7 out of 8 cases, the fractions of edited reads covered by the model was better than the 99th percentile of randomized outcome combinations and in 6 out of 8 cases and also superior to the 99.9th percentile. Overall, these results support that a large fraction of the gRNAs included in our library can edit their genomic targets in an efficient and predictable manner.

High throughput screening using the gRNA library

The gRNA library was cloned into a high-throughput co-selection base editing vector[23]. We performed pooled mutagenesis followed by bulk competition (Supplementary Figure 7) to identify mutations with significant fitness effects (Figure 2). 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[6,26].

After applying a stringent filtering threshold based on gRNA read count at the mutagenesis step (see methods), 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 showed high correlation across experimental time points (Supplementary Figure 8) and cluster by experimental step (Supplementary Figure 9), 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 at an estimated false positive rate of 5% (Figure 3A). GNEs are distributed evenly across the yeast genome (Figure 3B), suggesting no inherent bias against specific regions. An example of gRNA abundance variation through time for all gRNAs (both GNEs and NSGs) targeting *GLN4* is shown in Figure 3C.

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. The gRNA 33725 mutates a glycine at position 267 into either arginine or serine, and showed a dramatic drop in abundance in the large-scale experiment. To validate 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[27] in a heterozygous deletion background[28] (Supplementary figure 10A). 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[29]. 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 (Supplementary figure 10B), 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[30]. We found that mutating Gly 267 either to Arg or Ser was enough to cause protein loss of function (Figure 3D).

The five other sensitive sites identified in GLN4 by our screen were also clustered in regions with slow evolutionary rates. We found that one other GNE targeting residue D291 induced a highly deleterious mutation coupled with a neutral mutation as outcomes (D291E vs D291D, Supplementary Figure 11). We did not observe any discernible growth defect for the other GNE outcomes and as well as for the outcomes of 4 NSG targeting nearby amino acids. The other GNEs tested had markedly more positive scores than the one targeting G267, which would be consequent with a higher false positive rate close to the significance threshold. However, the case of the D291E/D291D pair, where a strong fitness effect is partially obscured by a neutral mutation produced by the other mutagenesis outcomes supports that sites of interest can be detected even close to the significance threshold. As we only tested two outcomes per gRNA, it is also possible that some of the abundance drops we measured were the result of mutations outside of our model, which are sometimes predicted to be more deleterious than the most likely mutations.

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 this using two recently published resources for variant effect prediction: Envision[2] and Mutfunc[31]. 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[32] 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 $\Delta\Delta G$ 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 (Supplementary Figure 12A). The structural modeling information had much lower coverage, covering at best around 12% of the most probable mutations (Supplementary Figure 12B). As expected, mutations generated by GNEs showed significantly lower SIFT scores and showed enrichment for strong effects predicted by SIFT and Envision (Figure 4). 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. Envision scores across the proteome show a high level of homogeneity, with most mutations having a score between 0.94 and 0.96 (Supplementary Figure 12C). According to the original Envision manuscript, this should be predictive of a small decrease in protein function. As such, the shifts in score distributions between GNEs and NSGs are more subtle but still support that GNE induced mutations are generally more likely to be deleterious as well (Supplementary Figure 13A).

Mutations with destabilizing effects as predicted by structural data also appeared to be enriched in GNEs predicted 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 (Supplementary Figure S13B-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[33], resulting in individual destabilizing mutations having a limited effects on fitness. As expected from known experimental data on mutagenesis outcomes[15], signal was usually stronger for the most probable C to G mutation.

Sensitive sites provide new biological insights

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 5A, source data in Supplementary tables 2, 3, and 4). We found deviations from random expectations in both C-to-G and C-to-T mutation ratios that drove the enrichment of several mutation combination. 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[34], making these substitutions highly deleterious. Furthermore, as Glycine residues are often important components of cofactor binding motifs (eg.: Phosphates)[35] this observation might reflect a tendency for GNEs to alter these sites.

As expected, there is a strong enrichment within GNEs for patterns that result in mutation to stop codons: both C-to-G patterns (Y to stop: 3 fold enrichment, $p=3.62x10^{-11}$, S to stop: 2.2 fold enrichment, p=0.0002) but only one C-to-T pattern was overrepresented significantly (W to stop, 4.6 fold enrichment, $p=6.23x10^{-15}$). Substitutions to stop codon in one outcome also drove enrichment in the other: for example, the link between Serine to Stop (C-to-G) appears to be the cause of the Serine to Leucine (C-to-T) overrepresentation. Both mutation pairs involving mutating a Tryptophan to a stop via a C-to-T mutation: this is not surprising, as the alternative mutations Tryptophan to Serine or Cysteine are also highly disruptive[34]. Changes between similar amino acids, which are expected to be tolerable, were also generally depleted in GNE (ex.: the Alanine to Glycine/Valine pair). Mutations in intronic sequences and putative non-functional peptides were also underrepresented, as were most patterns leading to silent

293 mutations (Figure 5A). These results show the power of this approach to discriminate important
294 functional sites from more mutation tolerative ones across the genome.
295
296 Interestingly, genes for which more than one GNE were detected were enriched for molecular
297 function terms linked to cofactor binding (Supplementary Table 5). This suggests that the GNEs
298 might indeed have a tendency to affect protein function through mechanisms other than protein
299 or interaction interface destabilization. These protein properties depend on many residues,
300 making them more robust to single amino acid substitutions, whereas cofactor binding may
301 depend specifically on a handful of residues, making these sites critical for function. Using the
302 Uniprot database[37], we also examined whether gRNAs that target annotated binding sites or
303 highly conserved motifs are more likely to affect fitness compared to other gRNAs targeting the
304 same set of genes. We found a 3.5 fold enrichment for GNEs directly affecting these sites
305 (49/188, $ratio^{GNE\ On}$=0.261, two-sided Fisher's exact test $p=3.54x10^{-14}$) or residues in a two
306 amino acid window around them (23/115, $ratio^{GNE\ near}$=0.167, two-sided Fisher's exact test
307 p=0.00048).
308
309 The precise targeting of our method also allows us to investigate amino acid residues with
310 known functional annotations such as post-translational modifications. We found no significant
311 enrichment for gRNAs mutating directly annotated PTMs ($ratio^{GNE\ PTM}$ = 19/1118, $ratio^{NSG\ PTM}$
312 243/15536, Fisher's exact test p=0.71). Most of these sites were phosphorylation sites (7),
313 metal coordinating residues (5) and ubiquitination sites (4). This is consistent with the
314 hypothesis that many PTM sites may have little functional importance[36] and thus mutations
315 affecting them should not be significantly enriched for strong fitness effects compared to other
316 possible mutations. The same was also observed for gRNAs mutating residues near known
317 PTMs that could disturb recognition sites ($ratio^{GNE\ nearPTM}$ = 130/1118, $ratio^{NSG\ nearPTM}$ =

1698/15536, Fisher's exact test p=0.43). As we did not specifically target PTMs, our sample size is small and it should be noted that statistical power regarding these observations is limited.

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[38,39], 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[40], which part of the only *RAP1* fragment essential for cell growth[41,42]. Because the available wild-type *RAP1* plasmid (see methods) does not complement gene deletion growth phenotype, we used a different strategy for validation that relied on CRISPR-mediated knock-in (see methods and Supplementary Figure 14). We tested the effect of several predicted GNE induced mutations in RAP1 targeting positions T486, A510, R523 and A540 (Figure 5B-C). We found that the predicted mutations at two of these positions, R523 and A540, were highly deleterious. While we could not validate that the two most likely mutations predicted to be caused by the GNE targeting T486 had a detectable fitness effect in these conditions, we found that phosphomimetic mutations at this position were lethal but most other amino acids were well tolerated. While we could validate that this gRNA indeed targeted a sensitive site, the outcomes predicted by our model did not have any detectable fitness effects. This showcases a limitation of our approach: the uncertainty in outcome prediction can complicate validation studies. 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. We therefore sought to examine what gRNA and target sequence features could influence mutagenesis efficiency. To do so, we focused on the subset of gRNAs with the potential to generate stop codons (stop codon generating gRNAs, SGGs) in essential genes (Figure 6A). As gRNAs in our library were designed to target the first 75% of the coding sequences, successful stop codon generation in this subset of genes should often lead to a lethal loss of function[13,22].

We found important variation in the ratio of GNE for the different types of SGGs (Figure 6B), with gRNAs targeting TGG (Trp) codons having the highest activity. This is in opposition to the general trend, as in general C to G mutation leading to stop codon formation had higher GNE ratios than the three other C-to-T alternatives. Overall, we observed significant GNE enrichment in SGGs which depend on the first C to G mutation to induce stop codon formation (Figure 6C). Multiple factors can explain the higher performance of TGG targeting gRNAs. First, as most of these sites have high co-editing risk scores because of the two consecutive cytosines, they might have increased editing rates due to processive co-editing events, increasing the chance of fitness effect detection. This phenomenon might also occur in non-SGG gRNAs (Supplementary Figure 15A). Second, we found a significant enrichment in GNEs for gRNAs targeting the non-coding strand, even after excluding SGGs (Figure 6D). This effect might be explained by the higher repair efficiency in the transcribed strand in yeast[43]. Furthermore, 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 when the mutated coding sequence is transcribed. In contrast, the targeted chromosomal strand appears to be much less important (Supplementary Figure 15B). The variation in GNE ratio observed between the different SGG target codons might also reflect *in vivo* DNA repair preferences that depend on sequence context, where different outcomes

might be favored depending on the target sequence. For example, the CA di-nucleotide might favor C to G mutations, which would explain the low GNE ratio of CAA (Gln) targeting SGGs and the higher than average GNE ratio of TCA (Ser) targeting SGGs.

Another parameter with a high impact on GNE enrichment in gRNA sets is the predicted melting temperature of the RNA-DNA duplex formed by the gRNA sequence and its target DNA sequence (Supplementary Figure 15C-D). Both SGG and non-SGG gRNAs with low values have a lower chance of being detected as having effects, while gRNAs with higher values are enriched for GNEs (Figure 6E). 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 any time point (Supplementary Figure 16). Furthermore, this effect is not caused by target position bias within target genes or a strong correlation between GC content and the targeted position (Supplementary Figure 17). Even if binding energy is strongly correlated with GC content, there is still significant variation within gRNA sets with the same %GC (Figure 6F).

Discussion

Using targeted deep sequencing and high throughput screening, we investigated whether the Target-AID base editor is amenable for genome-scale targeted mutagenesis studies. We show that a prediction model based on known Target-AID properties can be used to predict the major mutational outcome of editing, even if multiple editable nucleotides are present in the activity window. Using yeast essential genes as a test case, we then applied this approach on a larger scale and identified hundreds of gRNAs targeting sensitive residues that have significant effects on cellular fitness when mutated. We could then verify orthogonally the effects of mutational outcomes of GNE using classical genetics approaches and show that they tend to overlap with variants predicted to be deleterious. By focusing on a few highly relevant variant sets, we

highlighted the power and potential of our approach to generate new biological insights. We then 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 in specific genomic spaces.

In previously published methods such as TAM and CRISPR-X[18,19], 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 follow-up challenging. Using a base editor that channels mutational outcomes such as cytidine deaminase-uracil glycosylase inhibitor (UGI) fusion can address this problem[15] 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[44,45]. 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[15]. Recently, Sadhu, Bloom et al examined the effects of premature stop codons (PTC) in essential genes using a high throughput variant construction method that relied on homology directed repair using a mutated repair template[13]. They observed that a significant fraction of PTCs can be tolerated, but only within the last 30 codons of a protein. Outside this window, they found no link between

419 PTC tolerance and position within the coding sequence, something which we also did not
420 observe both for SGGs and non-SGG gRNAs (Supplementary Figure 17A-B).
421
422 We provide key empirical data on gRNA dependant parameters that can be used to optimize
423 base editing efficiency. Based on our results, selecting gRNAs with high binding energy to their
424 genomic targets and favoring those which target the non-coding strand can increase the chance
425 of high editing activity. Importantly, our observations differ from what has been reported for
426 Cas9-based genome editing. High gRNA RNA/DNA duplex binding has instead been associated
427 with lower mutagenesis efficiency[46]. Our data thus confirms the observation that parameters
428 associated with Cas9 editing cannot readily be transferred to base editors[47]. Furthermore, the
429 temperature at which experiments are performed might affect efficiency for certain gRNAs with
430 low gRNA-DNA duplex binding energy and should be considered when designing base editing
431 experiments in different organisms[15]. However, it remains to be confirmed whether the
432 enrichment for certain gRNA properties we observed are specific to Target-AID or will also be
433 transferable to other base editors as this may depend on the enzymatic properties of these
434 proteins. Acquiring large paired gRNA and mutagenesis outcome datasets similar to those
435 available for Cas9 genome editing[24] will allow for more refined models for rational base editing
436 activity prediction.
437
438 The field of base editing is rapidly evolving, with new tools being developed constantly. One of
439 the most recent additions to this fast-growing toolkit are engineered Cas9 enzymes with
440 broadened PAM specificities[48], which have already been shown to be compatible with base
441 editors. More flexible PAM requirements are especially useful for base editing applications, as
442 they increase the number of sites to be edited and also the number of potential gRNAs per site,
443 increasing the chances of choosing optimal properties and thus greater efficiency[25]. Our method

444 allows an experimental scale which bridges saturation mutagenesis methods and genome-wide
445 knock-out studies, alleviating the current trade-off between mutational diversity and the number
446 of targets genes to generate new biological insights.

447

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)[49] 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[50]. 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)[3,4] 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[51] 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[23]. 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)[52] and putative non-functional peptides[53] 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 Supplementary Figure 1. To assign a co-editing risk score to each gRNA, we defined four categories using the extended activity window sequence composition shown in Table 1.

Table 1: Sequence patterns of co-editing risk categories

Co-editing risk category	Very Low	Low	Moderate	High
Sequence patterns	NCDDDNN NDCDDNN NDDCDNN	NCDDCNN NDCDCNN NDDCCNN	NCDCNNN	NCCDNNN

N = any nucleotide, D = A or T or G

Library construction

The plasmids, oligonucleotides, and media used in this study are listed in as Supplementary tables 6, 7 and 8 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*)[54] 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.

Base editing time course and library preparation for deep sequencing

Cells were co-transformed with pKN1252 and the pDYSCKO plasmid bearing the gRNA of interest using the protocol described below for the large-scale experiment. Transformant plates were scraped by adding ~5 ml of sterile water, incubating a few minutes at room temperature, and then using a glass rake to resuspend colonies. The resuspended cells (one pool per guide) were used to inoculate two replicate cultures per guide. Cells went through the same induction protocol as for the large-scale experiment, but scaled down to a 24 deepwell plate (see Supplementary Figures 3 and 7). The volumes used were: 3 ml for the initial SC-UL+glucose culture, 4 ml for the SC-UL+glycerol step, 3 ml for the SC-UL+galactose step, and 3 ml for the liquid canavanine co-selection step. At the end of the galactose induction step, 100 µl of a 1/2000 dilution of each well was plated on SC-ULR+canavanine solid media to obtain editing survivor colonies. At the glycerol to galactose media switch, a ~1 OD pellet was sampled by spinning cells at 13 200 RPM and removing the media. Cell pellets were then stored at -80°C for

subsequent DNA extraction. The same method was used to sample ~1 OD at T=6 hours in galactose, ~2 OD at T=12 hours in galactose, and ~3 OD at the end of canavanine co-selection. Plates with selected colonies (edited at the CAN1 locus) were soaked in water and scraped, and 1.4 ml of the resulting cell suspension was sampled and stored.

Genomic DNA was extracted from cell pellets using a standard phenol-chloroform method from each sample[55] and quantified by NanoDrop (Thermofisher, Massachusetts, USA). For each sample, we aimed to sequence both the target edit site and the *CAN1* co-selection edit site. To multiplex the 240 samples in the same sequencing library, we used the row-column-plate-indexed PCR (RCP-PCR) approach[56]. Briefly, each target locus was amplified from genomic DNA and universal adapter sequences were added to each end of the amplicon. A 1/2500 dilution of the resulting product was then used as template with a set of 10 (rows) by 12 (column) primers used to index each sample in a second PCR reaction. All samples for the same locus were then pooled together and normalized according to electrophoresis gel band intensity and then purified using magnetic beads. A third and final PCR reaction on the purified pools was then used to add plate indexes and Illumina adapters: this reaction was performed in quadruplicate and the products from the four reactions were pooled together for purification. Sequencing was performed using the MiSeq Reagent Kit v3 on an Illumina MiSeq for 600 cycles (IBIS sequencing platform, Université Laval).

After sequencing, samples were demultiplexed using a custom python script with the reads being subdivided in four (plate barcode forward, row barcode, column barcode and plate barcode reverse). After demultiplexing, the forward and reverse reads were merged using the PANDA-Seq software[57]. Reads were then aligned to reference locus sequences using the Needle software from EMBOSS[58]. A custom script was then used to parse the alignments and

extract genotype information for each read. The sequencing reads for the base editing deep sequencing experiment were deposited on the NCBI SRA as accession number PRJNA552472.

Library transformation in yeast

Competent BY4741 (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) cells were first transformed with the pKN1252 (p315-GalL-Target-AID) plasmid using a standard lithium acetate method. 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[59]: 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 of 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[23] and is shown in Supplementary Figure 7. 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[60]. 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

590

591 Thermocycler protocol:

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

592

593 The resulting product was verified on a 2% agarose gel colored with Midori Green Advance
594 (Nippon Genetics, Japan) and then gel-extracted and purified using the FastGene Gel/PCR
595 Extraction Kit (Nippon Genetics, Japan). The purified products were used as the template for
596 the 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

597

598 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 are available on github (https://docker.pkg.github.com/Landrylab), and 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 [50] 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

663 streaked from the glycerol stock was used to prepare competent cells using the previously
664 described lithium acetate protocol[59]. To generate mutant alleles of the genes of interest, we
665 performed site-directed mutagenesis on the appropriate MoBY collection plasmid [27]. These
666 centromeric plasmids encode the yeast gene of interest under the control of their native
667 promoters and terminators. Mutagenesis reactions were performed with the following reaction
668 setup:

669

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

670

671 Thermocycler protocol:

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

672

673 After amplification, the mutagenesis product was digested with DpnI for 2 hours at 37°C and 5 ul
674 was transformed in *E. coli* strain BW23474 (F-, Δ*(argF-lac)169*, Δ*uidA4::pir-116*, *recA1*,
675 *rpoS396(Am)*, *endA9(del-ins)::FRT*, *rph-1*, *hsdR514*, *rob-1*, *creC510*)[61]. Transformants were
676 plated on 2YT+Kan+Chlo and grown at 37°C overnight. Plasmid DNA was then isolated from

677 clones and sent for Sanger sequencing (CHUL sequencing platform, Université Laval, Québec
678 City, Canada) to confirm mutagenesis success.

679 Competent cells of target genes were transformed with the appropriate mutant plasmids as well
680 a the original plasmid bearing the wild-type gene and the empty vector[62], and transformants
681 were selected by plating on SC-U (MSG). Multiple independent colonies per transformation
682 were then put on sporulation media until sporulation could be confirmed by microscopy. For
683 tetrad dissection, cells were resuspended in 100ul 20T zymolyase (200mg/ml dilution in water)
684 and incubated for 20 minutes at room temperature. Cells were then centrifuged and
685 resuspended in 50ul 1M sorbitol before being streaked on a level YPD plate. All dissections
686 were performed using a Singer SporePlay microscope (Singer Instruments, UK). Plate pictures
687 were taken after five days incubation at room temperature except for the RAP1 plasmid
688 complementation test for which the picture was taken after three days. Pictures are shown in
689 Supplementary Image File 1.

690

691 **Strain construction for confirmations in *RAP1***

692 Because the MoBY collection plasmid for RAP1 cannot fully complement the gene deletion
693 (Supplementary image file 1), we instead performed confirmations by engineering mutations a
694 diploid strain to create heterozygous mutants. *RAP1* was first tagged with a modified version of
695 fragment DHFR F[1,2] (the first half) of the mDHFR enzyme[63]. The mDHFR[1,2]-FLAG cassette
696 was amplified using gene-specific primers and previously described reaction parameters[63]. Cells
697 were transformed with the cassette using the previously described transformation protocol and
698 were plated on YPD+Nourseothricine (YPD+Nat in Media table). Positive clones were identified
699 by colony PCR and successful fragment fusion was confirmed by Sanger sequencing (CHUL
700 sequencing platform). We then mated the confirmed clones with strain Y8205 (*Matα*

can1::STE2pr-his5 lyp1::STE3prLEU2 ∆ura3 ∆his3 ∆leu2, Kindly gifted by Charlie Boone) by inoculating a 4ml YPD culture with overnight starter cultures of both strains and letting the culture grow overnight. Cells were then streaked on YPD+Nat and diploid cells were identified by colony PCR using mating type diagnosis primers[64].

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

CRISPR-Cas9 mediated Knock-in of targeted mutations

Mutant alleles of target genes were amplified in two fragments using template DNA from the haploid tagged strain (See Supplementary figure 14). The two fragments bearing mutations were then fused together by a second PCR round to form the final donor DNA. This DNA was then co-transformed with a plasmid bearing Cas9 and a gRNA targeting the URA3 cassette for HDR mediated editing using a standard protocol[65]. Clones were then screened by PCR to verify donor DNA and mutation integration at the target locus. The targeted region of *RAP1* was then Sanger sequenced (CHUL sequencing platform, Univesité Laval, Québec City, Canada) to confirm the 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 File 2).

Evolutionary rate measurements and protein variant abundance

Evolutionary rates were calculated using the Rate4site software[66] using multiple sequence alignments and phylogenies from PhylomeDB V4[67] 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[68] 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[69]. 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[70], which uses values taken from Sugimoto et al[71].

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[31]. 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[2] 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 Supplementary Figure 12.

We downloaded the Uniprot database for yeast genes (query: uniprot-proteome_UP000002311) with annotations covering the following properties: Metal binding, Nucleotide binding, Site, DNA binding, Calcium binding, Binding site, Active site, Motif. We found that 6295 gRNAs targeted genes which have annotations in Uniprot, of which 519 were GNEs ($ratio_{All}$=0.0749). Statistical enrichments were calculated using this set of gRNAs as the reference population. Gene enrichments were performed using the PANTHER gene list analysis tool[72]. 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.

Data Availability

All raw sequencing data has been deposited on the NCBI as accession number PRJNA552472. The gRNA screen scores, predicted mutation outcomes, mutation effect predictors scores, as well as other relevant annotations are presented in Supplementary Dataset 1. Source image files for the tetrad dissections are presented as Supplementary Image 1 and 2.

Code Availability

The custom Python scripts used to analyze the are available on github (https://github.com/landrylaboratory), and packages and software us ed are presented in Supplementary table 9.

Acknowledgments

This work was supported by the Canadian Institutes of Health Research Foundation grant 387697 to CR.L., 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

1. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–7 (2014).

2. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst.* (2018). doi:10.1016/j.cels.2017.11.003

3. Winzeler, E. A. *et al.* Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science (80-.).* **285**, 901–906 (1999).

4. Giaever, G. *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387–391 (2002).

5. C. elegans Deletion Mutant Consortium, T. C. elegans D. M. Large-Scale Screening for Targeted Knockouts in the Caenorhabditis elegans Genome. *G3; Genes|Genomes|Genetics* **2**, 1415–1425 (2012).

6. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science (80-.).* **343**, 84–87 (2014).

7. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–83 (2013).

8. Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32**, 347–55 (2014).

9. Smith, J. D. *et al.* Quantitative CRISPR interference screens in yeast identify chemical-genetic interactions and new rules for guide RNA design. *Genome Biol.* **17**, 45 (2016).

10. Sharon, E. *et al.* Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell* (2018). doi:10.1016/j.cell.2018.08.057

11. Bao, Z. *et al.* Genome-scale engineering of Saccharomyces cerevisiae with single-nucleotide precision. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4132

12. Roy, K. R. *et al.* Multiplexed precision genome editing with trackable genomic barcodes in yeast. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4137

13. Sadhu, M. J. *et al.* Highly parallel genome variant engineering with CRISPR-Cas9. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0087-y

14. Guo, X. *et al.* High-throughput creation and functional profiling of DNA sequence variant libraries using CRISPR-Cas9 in yeast. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4147

15. Nishida, K. *et al.* Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science (80-.).* **353**, 553–563 (2016).

16. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).

17. Rees, H. A. & Liu, D. R. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.* **19**, 770–788 (2018).

18. Ma, Y. *et al.* Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. *Nat. Methods* **13**, 1029–1035 (2016).

19. Hess, G. T. *et al.* Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods* (2016). doi:10.1038/nmeth.4038

20. Giaever, G. *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**, 387–391 (2002).

21. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).

22. Michel, A. H. *et al.* Functional mapping of yeast genomes by saturated transposition. *Elife* **6**, (2017).

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

24. Allen, F. *et al.* Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.* (2019). doi:10.1038/nbt.4317

25. Dandage, R., Després, P. C., Yachie, N. & Landry, C. R. beditor: A Computational Workflow for Designing Libraries of Guide RNAs for CRISPR-Mediated Base Editing. *Genetics* **212**, 377–385 (2019).

26. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).

27. Ho, C. H. *et al.* A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. *Nat. Biotechnol.* **27**, 369–377 (2009).

28. Giaever, G. *et al.* Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat. Genet.* **21**, 278–83 (1999).

29. Eriani, G., Delarue, M., Poch, O., Gangloff, J. & Moras, D. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature* **347**, 203–206 (1990).

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

31. Wagih, O. *et al.* A resource of variant effect predictions of single nucleotide variants in model organisms. *Mol. Syst. Biol.* **14**, e8430 (2018).

32. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* (2003).

33. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* **6**, 678–687 (2005).

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

873 35. Copley, R. R. & Barton, G. J. A Structural Analysis of Phosphate and Sulphate Binding
874 Sites in Proteins. *J. Mol. Biol.* **242**, 321–329 (1994).

875 36. Landry, C. R., Levy, E. D. & Michnick, S. W. Weak functional constraints on
876 phosphoproteomes. *Trends Genet.* **25**, 193–7 (2009).

877 37. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* (2019).
878 doi:10.1093/nar/gky1049

879 38. Albuquerque, C. P. *et al.* A multidimensional chromatography technology for in-depth
880 phosphoproteome analysis. *Mol. Cell. Proteomics* **7**, 1389–96 (2008).

881 39. Holt, L. J. *et al.* Global analysis of Cdk1 substrate phosphorylation sites provides insights
882 into evolution. *Science* **325**, 1682–6 (2009).

883 40. Konig, P., Giraldo, R., Chapman, L. & Rhodes, D. The crystal structure of the DNA-
884 binding domain of yeast RAP1 in complex with telomeric DNA. *Cell* **85**, 125–36 (1996).

885 41. Graham, I. R., Haw, R. A., Spink, K. G., Halden, K. A. & Chambers, A. In vivo analysis of
886 functional regions within yeast Rap1p. *Mol. Cell. Biol.* **19**, 7481–90 (1999).

887 42. Wu, A. C. K. *et al.* Repression of Divergent Noncoding Transcription by a Sequence-
888 Specific Transcription Factor. *Mol. Cell* **72**, 942-954.e7 (2018).

889 43. Reis, A. M. C. *et al.* Targeted detection of in vivo endogenous DNA base damage reveals
890 preferential base excision repair in the transcribed strand. *Nucleic Acids Res.* **40**, 206–
891 219 (2012).

892 44. Jin, S. *et al.* Cytosine, but not adenine, base editors induce genome-wide off-target
893 mutations in rice. *Science* eaaw7166 (2019). doi:10.1126/science.aaw7166

894 45. Zuo, E. *et al.* Cytosine base editor generates substantial off-target single-nucleotide
895 variants in mouse embryos. *Science (80-.).* eaav9973 (2019).
896 doi:10.1126/SCIENCE.AAV9973

897 46. Wong, N., Liu, W. & Wang, X. WU-CRISPR: characteristics of functional guide RNAs for
898 the CRISPR/Cas9 system. *Genome Biol.* **16**, 218 (2015).

899 47. Kim, D. *et al.* Genome-wide target specificities of CRISPR RNA-guided programmable
900 deaminases. *Nat. Biotechnol.* (2017). doi:10.1038/nbt.3852

901 48. Nishimasu, H. *et al.* Engineered CRISPR-Cas9 nuclease with expanded targeting space.
902 *Science* eaas9129 (2018). doi:10.1126/science.aas9129

903 49. Dicarlo, J. E. *et al.* Genome engineering in Saccharomyces cerevisiae using CRISPR-
904 Cas systems. *Nucleic Acids Res.* **41**, 4336–4343 (2013).

905 50. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient
906 alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

907 51. Qian, W., Ma, D., Xiao, C., Wang, Z. & Zhang, J. The Genomic Landscape and
908 Evolutionary Resolution of Antagonistic Pleiotropy in Yeast. *Cell Rep.* **2**, 1399–1410
909 (2012).

910 52. Grate, L. & Ares, M. Searching yeast intron data at Ares lab web site. *Methods Enzymol.*
911 (2002). doi:10.1016/S0076-6879(02)50975-7

912 53. Smith, J. E. *et al.* Translation of Small Open Reading Frames within Unannotated RNA
913 Transcripts in Saccharomyces cerevisiae. *Cell Rep.* **7**, 1858–1866 (2014).

914 54. Casadaban, M. J. & Cohen, S. N. Analysis of gene control signals by DNA fusion and
915 cloning in Escherichia coli. *J. Mol. Biol.* **138**, 179–207 (1980).

916 55. Amberg, D. C., Burke, D. J. & Strathern, J. N. *Methods in Yeast Genetics: A Cold Spring*
917 *Harbor Laboratory Course Manual, 2005 Edition*. *A Cold Spring Harbor Laboratory*
918 *Course Manual* (2005).

919 56. Yachie, N. *et al.* Pooled-matrix protein interaction screens using Barcode Fusion
920 Genetics. *Mol. Syst. Biol.* (2016). doi:10.15252/msb.20156660

921 57. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D.
922 PANDAseq: Paired-end assembler for illumina sequences. *BMC Bioinformatics* (2012).
923 doi:10.1186/1471-2105-13-31

924 58. Rice, P., Longden, L. & Bleasby, A. EMBOSS: The European Molecular Biology Open
925 Software Suite. *Trends in Genetics* (2000). doi:10.1016/S0168-9525(00)02024-2

926 59. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS
927 carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).

928 60. Yachie, N. *et al.* Pooled-matrix protein interaction screens using Barcode Fusion
929 Genetics. *Mol. Syst. Biol.* **12**, 863 (2016).

930 61. Haldimann, A. *et al.* Altered recognition mutants of the response regulator PhoB: a new
931 genetic strategy for studying protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.*
932 **93**, 14361–6 (1996).

933 62. Zhao, L. *et al.* A genome-wide imaging-based screening to identify genes involved in
934 synphilin-1 inclusion formation in Saccharomyces cerevisiae. *Sci. Rep.* **6**, 30134 (2016).

935 63. Tarassov, K. *et al.* An in vivo map of the yeast protein interactome. *Science* **320**, 1465–
936 70 (2008).

937 64. Huxley, C., Green, E. D. & Dunham, I. Rapid assessment of S. cerevisiae mating type by
938 PCR. *Trends Genet.* **6**, 236 (1990).

939 65. Ryan, O. W., Poddar, S. & Cate, J. H. D. Crispr–cas9 genome engineering in
940 Saccharomyces cerevisiae cells. *Cold Spring Harb. Protoc.* **2016**, 525–533 (2016).

941 66. Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. Comparison of site-specific rate-inference
942 methods for protein sequences: Empirical Bayesian methods are superior. *Mol. Biol.*
943 *Evol.* (2004). doi:10.1093/molbev/msh194

944 67. Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M. & Gabaldón,
945 T. PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome.
946 *Nucleic Acids Res.* (2014). doi:10.1093/nar/gkt1177

947 68. Pearson, W. R., Wood, T., Zhang, Z. & Miller, W. Comparison of DNA Sequences with
948 Protein Sequences. *Genomics* **46**, 24–36 (1997).

949 69. Cherry, J. M. *et al.* Saccharomyces Genome Database: The genomics resource of
950 budding yeast. *Nucleic Acids Res.* (2012). doi:10.1093/nar/gkr1029

951 70. Cock, P. J. A. *et al.* Biopython: Freely available Python tools for computational molecular

952 biology and bioinformatics. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp163

953 71. Sugimoto, N. *et al.* Thermodynamic parameters to predict stability of RNA/DNA hybrid
954 duplexes. *Biochemistry* **34**, 11211–6 (1995).

955 72. Mi, H. *et al.* Protocol Update for large-scale genome and gene function analysis with the
956 PANTHER classification system (v.14.0). *Nat. Protoc.* **14**, 703–721 (2019).

957 73. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**,
958 W382–W388 (2005).

959

960

961 FIGURE LEGENDS
962

Figure 1 A simple parsimonious model predicts the most probable outcomes of Target-AID mutagenesis. A) gRNAs included in the time course base editing experiment had diverse C content profiles in the Target-AID activity window. Nucleotides are color coded: guanines are purple, thymines are red, adenines are green and cytosines are blue. **B)** Overall fraction of edited reads for all target sites rate along timepoints in the experiment: T0 (start of induction), T6 (mid induction), T12 (end of induction). The solid time point represents surviving cells plated after galactose induction, while the liquid time point represents the cell population after canavanine co-selection. Amplification of the *ERO1* target site from the liquid recovery time points was unsuccessful (shown in grey), and as such the solid recovery time point was used instead for the other analysis steps. **C)** Fraction of genotypes with different numbers of edited nucleotides in the Target-AID activity window after co-selection for each locus. Values represents the fraction of reads with either one, two or three edits compared to the total fraction of reads that were edited. **D)** Editing outcome type for all sites with a total editing rate greater than one percent after co-selection (n=30 cytosines across all targeted sites). The C to G/T distribution represents the sum of editing that resulted in a C to G or C to T mutation. Position-wise editing rates and outcome are shown in Supplementary Figures 5 and 6. **E)** Agreement between the predicted nucleotide total editing rank in the model used to predict mutagenesis outcomes in the large-scale experiment and the deep sequencing data (n=28 sites, 10 gRNAs). The gRNAs targeting *ADE1* and *SES1* were respectively excluded from the analysis because there is only one editable site in the activity window and total editing rate was too low. **F)** Edited read coverage of the mutation outcome prediction model and the 99th percentile of edited allele combinations (n=4 genotypes in both cases) for the gRNAs with editing activity included in the large-scale experiment.

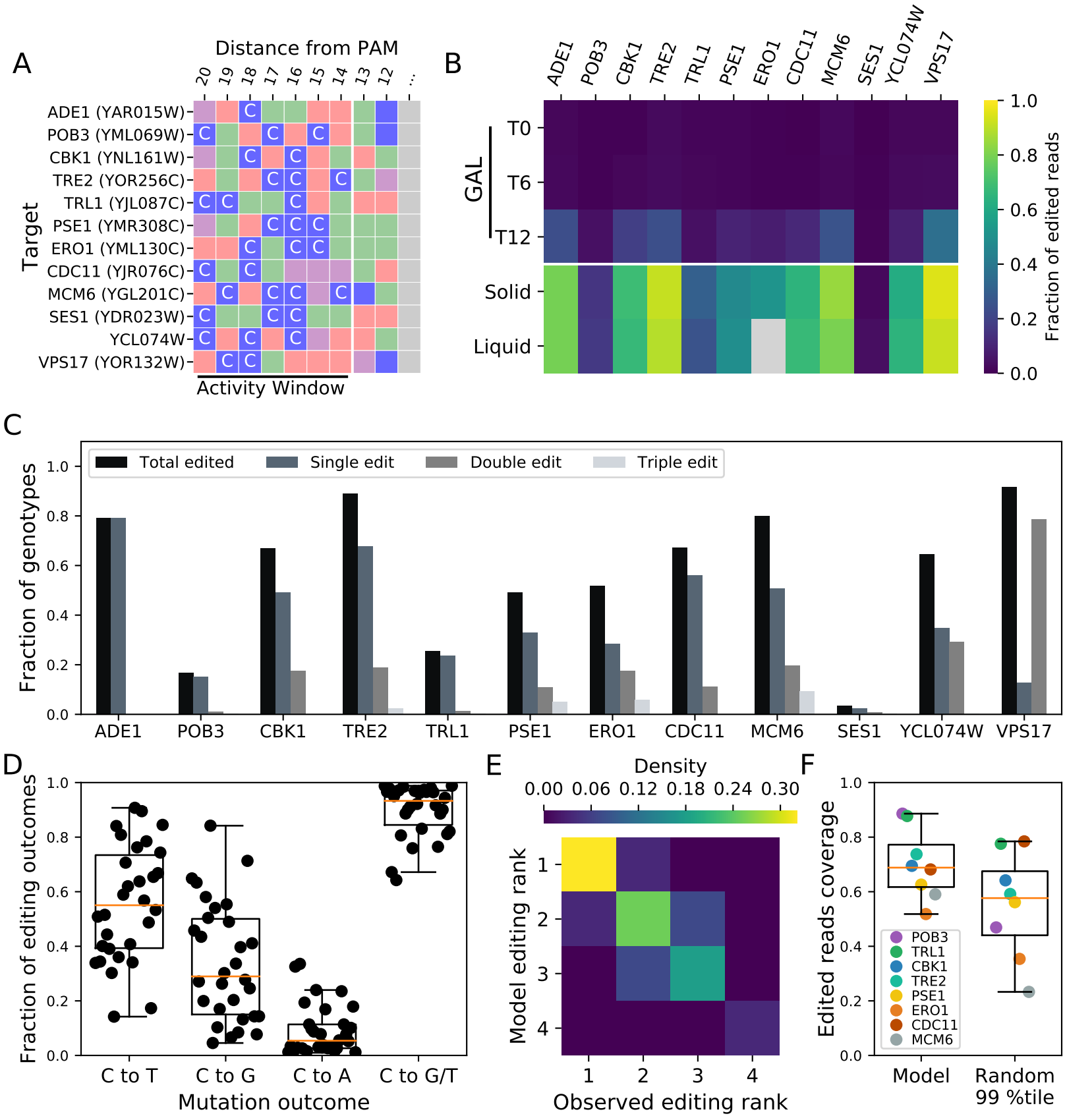
985 **Figure 2 A gRNA library for systematic perturbation of essential genes using the Target-AID base**
986 **editor.** Essential genes (ex.: *E.G.1*) were scanned for sites appropriate for Target-AID mutagenesis.
987 Mutational outcomes include silent (grey triangle), missense (black triangle) mutations, as well as stop
988 codons (*). DNA fragments corresponding to the gRNA sequences were synthesized as an
989 oligonucleotide pool and cloned into a co-selection base editing vector. Using gRNAs as molecular
990 barcodes, the abundance of cell subpopulations bearing mutations is then measured after mutagenesis
991 and bulk competition. Mutations with fitness effects are inferred from reductions in the relative gRNA
992 abundances.
993

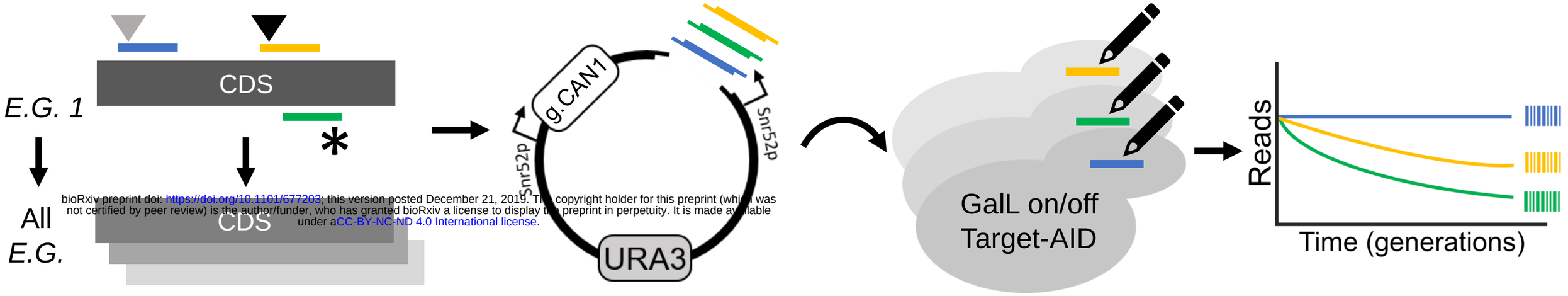
Figure 3. High-throughput forward mutagenesis by Target-AID base editing identifies sensitive sites across the yeast genome. A) Cumulative distribution of z-scores of the log2 fold-change in gRNA abundance between mutagenesis and the end of the bulk competition experiment averaged between replicates. A 5% false positive threshold was calculated by fitting a distribution to the 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. **B)** 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. **C)** Decline in gRNA abundance (on a log scale) between timepoints after mutagenesis for gRNAs targeting GLN4, a tRNA synthetase. Median gRNA 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. **D)** Mutagenesis of GLN4-G267 confirms its essential role for protein function. 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.

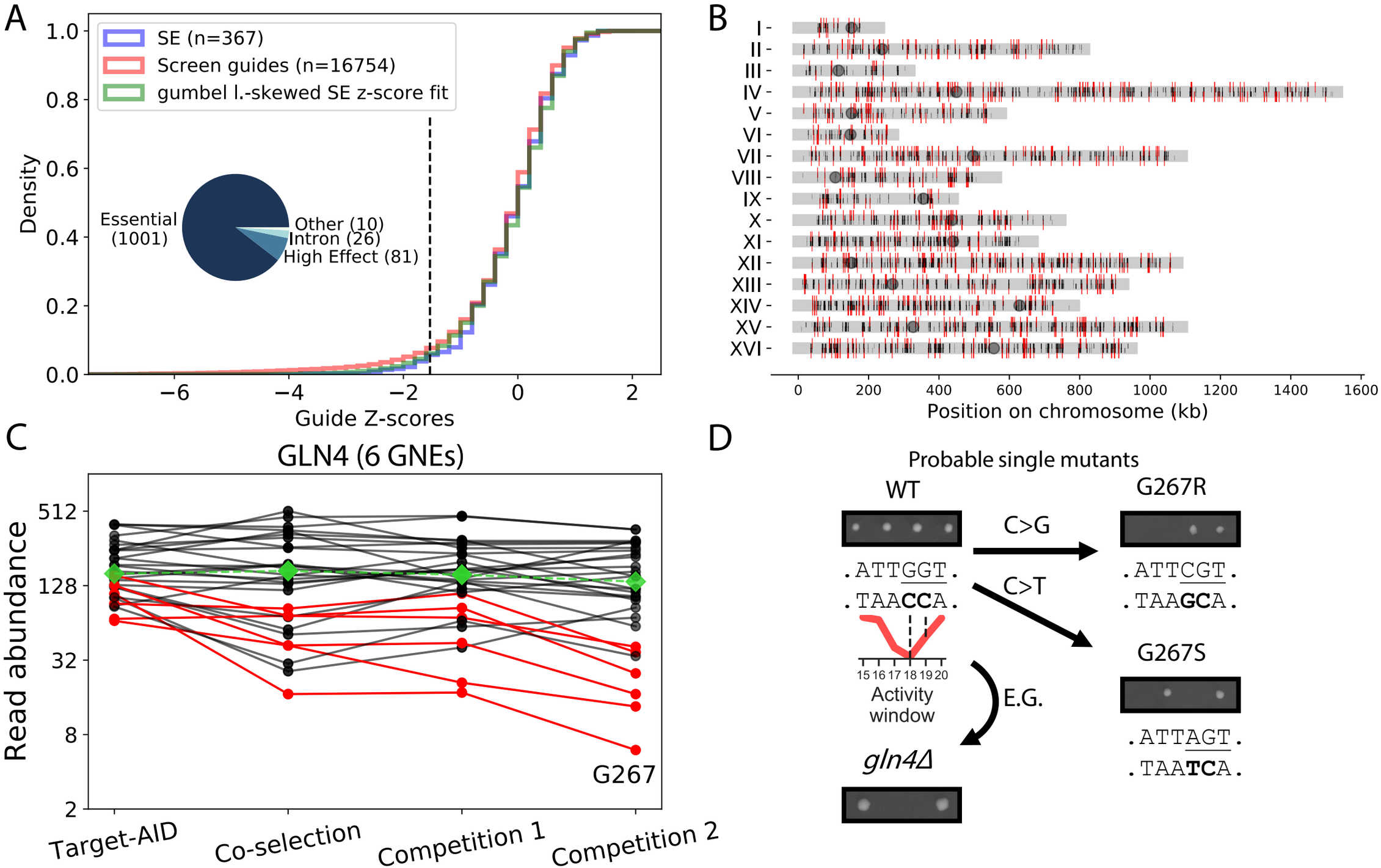
Figure 4: GNE induced mutations are enriched in predicted deleterious effects A) SIFT score distributions for the most likely induced mutations of both GNEs (blue) and NSGs (red). The thresholds for the categories used in the enrichment calculations in **B)** are shown as black dotted lines. SIFT scores 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[32] but might be permissive considering the number of mutations tested in our experiment (n= 895, 12394, 704, 8520, 643, 7396, 508, 5682). All GNE vs NSG score comparisons are significant (Welch's t-test p-values: $1.19x10^{-24}$, $3.01x10^{-24}$, $9.00x10^{-12}$, $1.55x10^{-12}$). 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. $\Delta\Delta$G), protein folding based on homology models (Model $\Delta\Delta$G) and protein-protein interaction interface stability based on structure data (Inter. $\Delta\Delta$G). The raw values used to calculate ratios are shown in Supplementary table 1. The predictions based on conservation and experimental data are grouped under 'Predictors' and those based on the computational analysis of protein structures and complexes under 'Structural'.

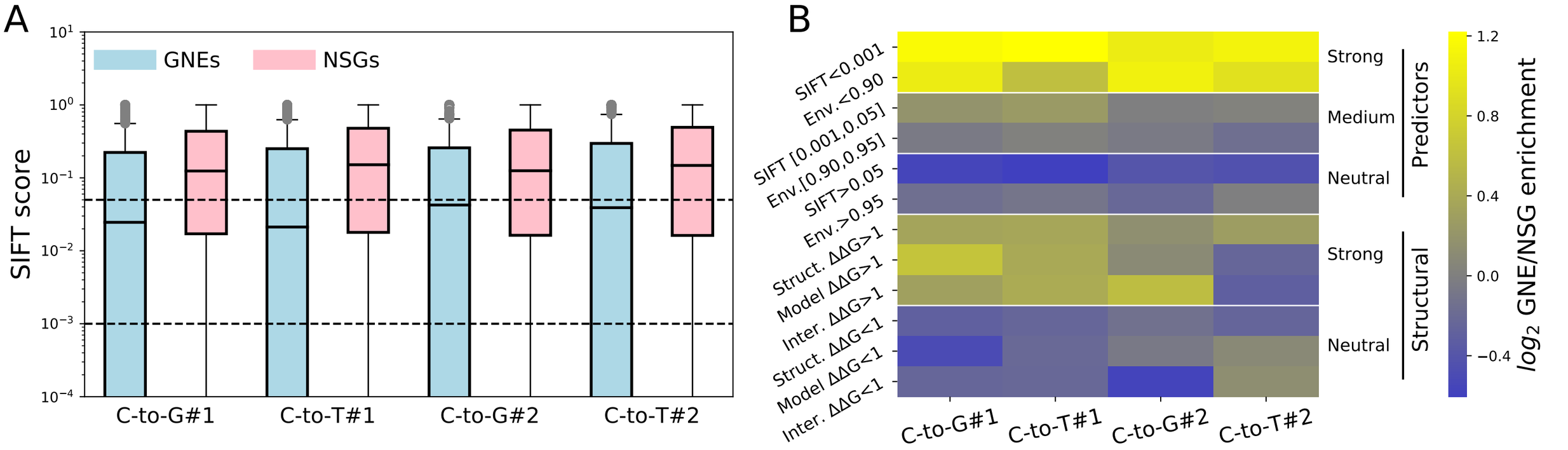
Figure 5 GNE mutations are enriched for specific amino acid substitution patterns and identify critical sites for protein function. A) Fold depletion and enrichment volcano plots for the most probable mutations induced by GNEs in the screen. Enrichment and depletion values were calculated by comparing 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.

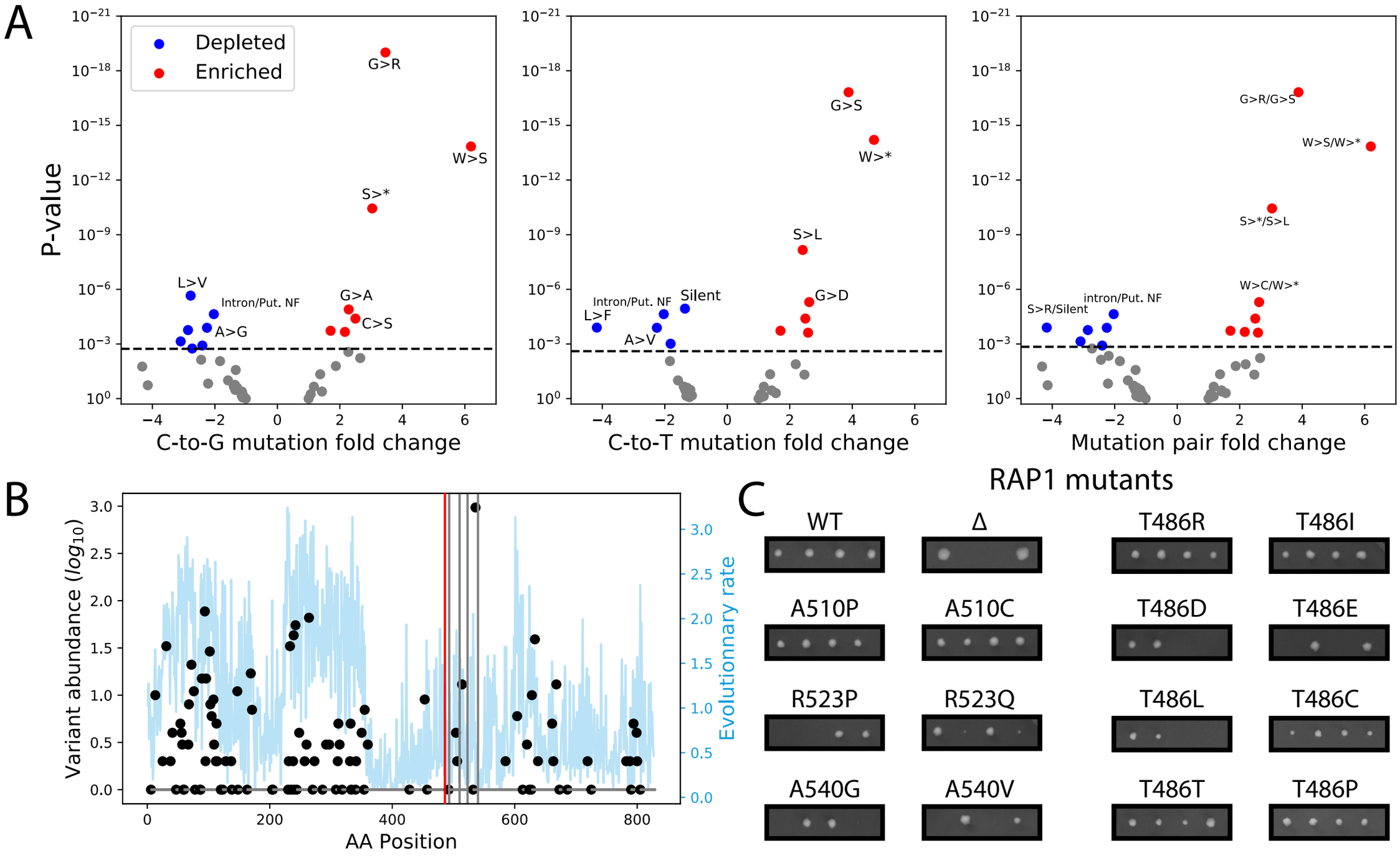
Figure 6 gRNA and target properties affect mutagenesis efficiency. A) Since Target-AID can generate both C to G and C to T mutations, many codons can be targeted to create premature stop codons. The TGG (W) codon is the only one targeted on the non-coding strand as ACC. **B)** GNE ratio for SGGs targeting different codons in essential genes, split by co-editing risk categories. **C)** 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. **D)** Cumulative z-score density of gRNAs that do not generate stop codons targeting either the coding or non-coding strand. **E)** SGG and non-SGG GNE enrichment compared to the expected GNE ratio for different melting temperature ranges. **F)** gRNA/DNA duplex melting temperature as a function of gRNA GC content for all gRNAs for which fitness effects were measured. The higher and lower efficiency thresholds are based on the enrichments shown in panel E.

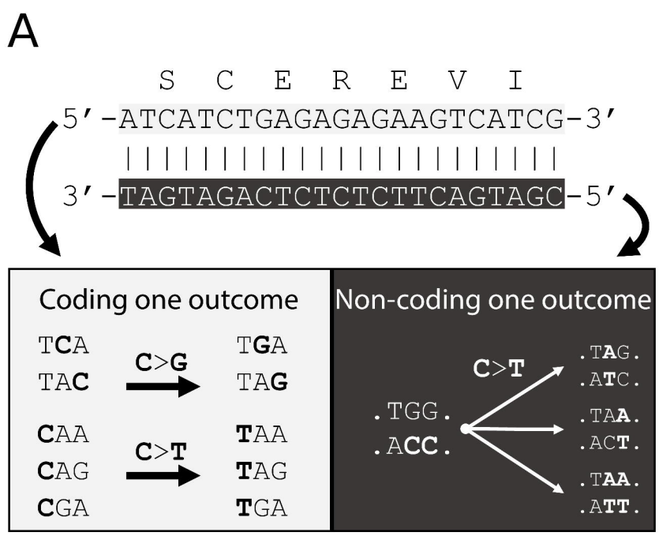

A
Distance from PAM
20
19
18
17
16
15
14
13
12
...
Target
ADE1 (YAR015W)
POB3 (YML069W)
CBK1 (YNL161W)
TRE2 (YOR256C)
TRL1 (YJL087C)
PSE1 (YMR308C)
ERO1 (YML130C)
CDC11 (YJR076C)
MCM6 (YGL201C)
SES1 (YDR023W)
YCL074W
VPS17 (YOR132W)
Activity Window
B
ADE1
POB3
CBK1
TRE2
TRL1
PSE1
ERO1
CDC11
MCM6
SES1
YCL074W
VPS17
GAL
T0
T6
T12
Solid
Liquid
Fraction of edited reads
1.0
0.8
0.6
0.4
0.2
0.0
C
Total edited
Single edit
Double edit
Triple edit
Fraction of genotypes
1.0
0.8
0.6
0.4
0.2
0.0
ADE1
POB3
CBK1
TRE2
TRL1
PSE1
ERO1
CDC11
MCM6
SES1
YCL074W
VPS17
D
Fraction of editing outcomes
1.0
0.8
0.6
0.4
0.2
0.0
C to T
C to G
C to A
C to G/T
Mutation outcome
E
Density
0.00 0.06 0.12 0.18 0.24 0.30
Model editing rank
1
2
3
4
1
2
3
4
Observed editing rank
F
Edited reads coverage
1.0
0.8
0.6
0.4
0.2
0.0
POB3
TRL1
CBK1
TRE2
PSE1
ERO1
CDC11
MCM6
Model
Random
99 %tile

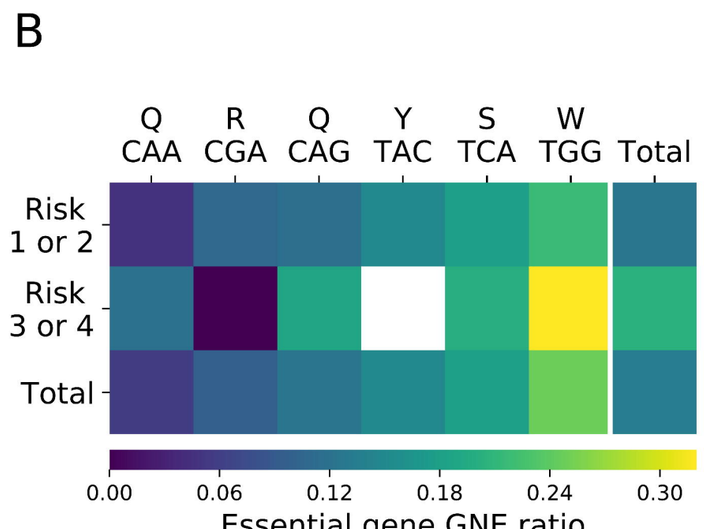


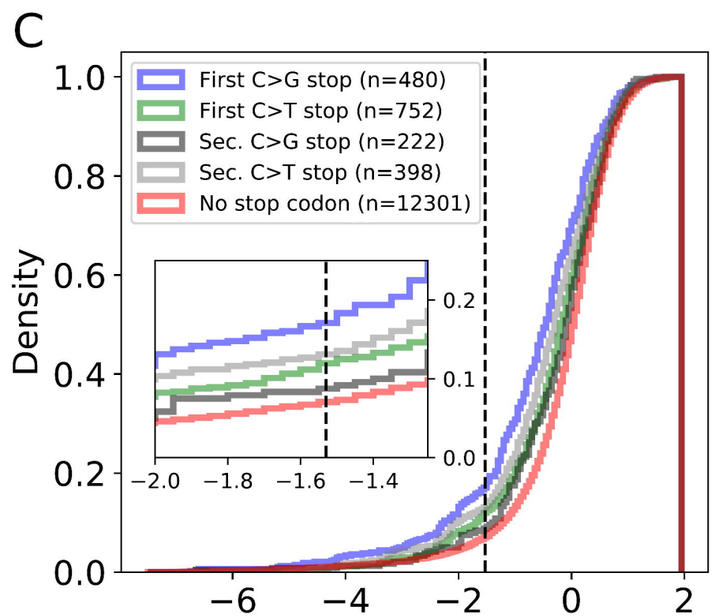

A
SE (n=367)
Screen guides (n=16754)
gumbel l.-skewed SE z-score fit
Density
Guide Z-scores
Essential (1001)
Other (10)
Intron (26)
High Effect (81)
B
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
XIII
XIV
XV
XVI
Position on chromosome (kb)
C
GLN4 (6 GNEs)
Read abundance
G267
Target-AID
Co-selection
Competition 1
Competition 2
D
Probable single mutants
WT
G267R
C>G
C>T
.ATTGGT.
.TAACCA.
.ATTCGT.
.TAAGCA.
15 16 17 18 19 20
Activity window
E.G.
G267S
gln4Δ
.ATTAGT.
.TAATCA.

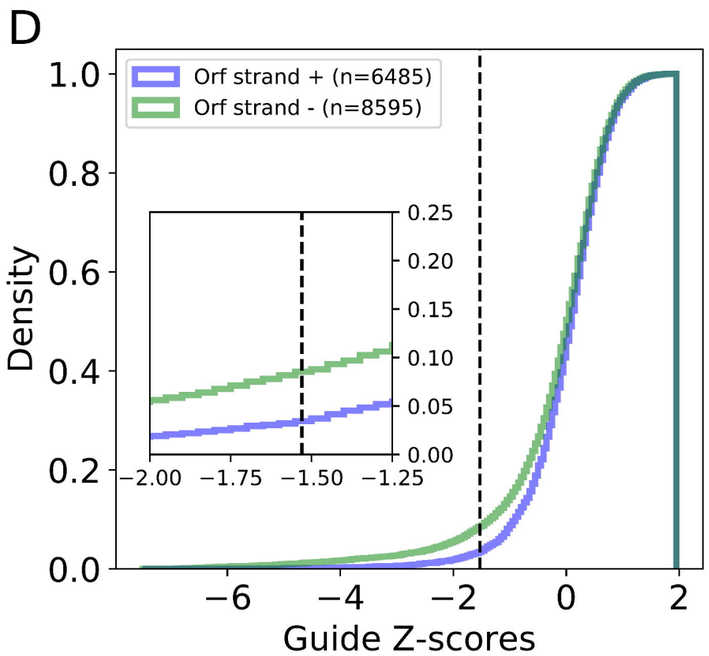

A
GNEs
NSGs
SIFT score
10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
C-to-G#1
C-to-T#1
C-to-G#2
C-to-T#2
B
SIFT<0.001
Env.<0.90
SIFT [0.001,0.05]
Env.[0.90,0.95]
SIFT>0.05
Env.>0.95
Struct. ΔΔG>1
Model ΔΔG>1
Inter. ΔΔG>1
Struct. ΔΔG<1
Model ΔΔG<1
Inter. ΔΔG<1
C-to-G#1
C-to-T#1
C-to-G#2
C-to-T#2
Strong
Medium
Neutral
Predictors
Strong
Neutral
Structural
1.2
0.8
0.4
0.0
−0.4
log_2 GNE/NSG enrichment

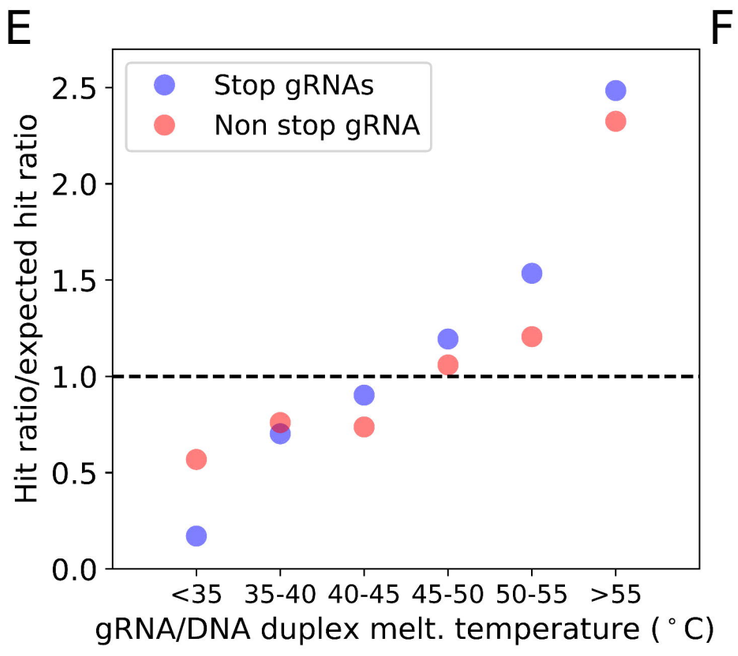

A
Depleted
Enriched
P-value
G>R
W>S
S>*
L>V
Intron/Put. NF
A>G
G>A
C>S
C-to-G mutation fold change
G>S
W>*
S>L
Intron/Put. NF
Silent
L>F
A>V
G>D
C-to-T mutation fold change
G>R/G>S
W>S/W>*
S>*/S>L
W>C/W>*
S>R/Silent
intron/Put. NF
Mutation pair fold change
B
Variant abundance (log10)
Evolutionnary rate
AA Position
C
RAP1 mutants
WT
Δ
T486R
T486I
A510P
A510C
T486D
T486E
R523P
R523Q
T486L
T486C
A540G
A540V
T486T
T486P

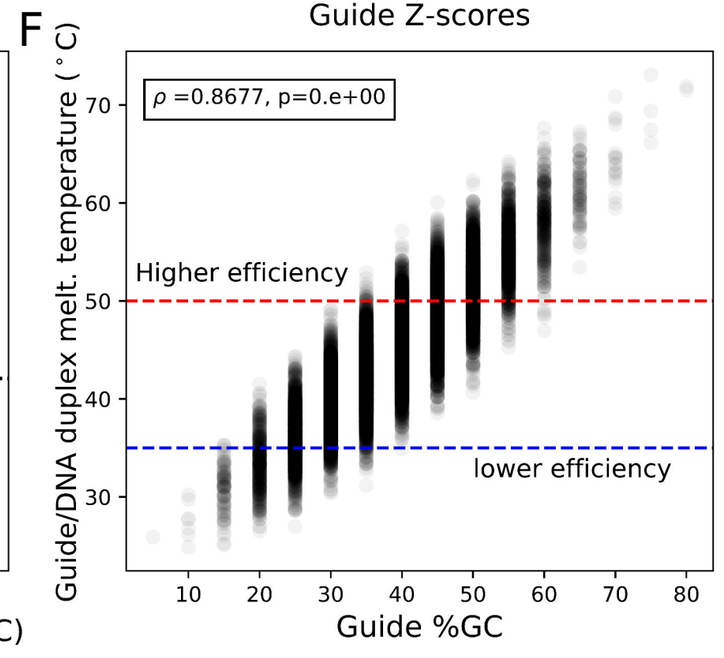

A
S C E R E V I
5′-ATCATCTGAGAGAGAAGTCATCG-3′
3′-TAGTAGACTCTCTCTTCAGTAGC-5′
Coding one outcome
TCA
TAC
C>G
TGA
TAG
CAA
CAG
CGA
C>T
TAA
TAG
TGA
Non-coding one outcome
C>T
.TGG.
.ACC.
.TAG.
.ATC.
.TAA.
.ACT.
.TAA.
.ATT.


B
Q CAA
R CGA
Q CAG
Y TAC
S TCA
W TGG
Total
Risk 1 or 2
Risk 3 or 4
Total
0.00
0.06
0.12
0.18
0.24
0.30
Essential gene GNE ratio


C
First C>G stop (n=480)
First C>T stop (n=752)
Sec. C>G stop (n=222)
Sec. C>T stop (n=398)
No stop codon (n=12301)
Density
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
0.2
0.1
0.0
−2.0
−1.8
−1.6
−1.4


D
Orf strand + (n=6485)
Orf strand - (n=8595)
Density
1.0
0.8
0.6
0.4
0.2
0.0
0.25
0.20
0.15
0.10
0.05
0.00
−2.00
−1.75
−1.50
−1.25
−6
−4
−2
0
2
Guide Z-scores


E
Stop gRNAs
Non stop gRNA
Hit ratio/expected hit ratio
2.5
2.0
1.5
1.0
0.5
0.0
<35
35-40
40-45
45-50
50-55
>55
gRNA/DNA duplex melt. temperature (˚C)


F
Guide Z-scores
ρ =0.8677, p=0.e+00
Higher efficiency
lower efficiency
Guide/DNA duplex melt. temperature (˚C)
70
60
50
40
30
10
20
30
40
50
60
70
80
Guide %GC