





# Perturbing proteomes at single residue resolution using base editing

Philippe C. Després <sup>1,2,3,4</sup>, Alexandre K. Dubé <sup>1,2,3,4,5</sup>, Motoaki Seki<sup>6</sup>, Nozomu Yachie <sup>6,7,8</sup>✉ & Christian R. Landry <sup>1,2,3,4,5</sup>✉

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 the Target-AID base editor to systematically modify genes genome-wide by targeting yeast essential genes. We mutate around 17,000 individual sites in parallel across more than 1500 genes. We identify over 700 sites at which mutations have a significant impact on fitness. Using previously determined and preferred Target-AID mutational outcomes, we find that gRNAs with significant effects on fitness are enriched in variants predicted to be deleterious based on residue conservation and predicted protein destabilization. We identify key features influencing 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.

<sup>1</sup>Département de Biochimie, Microbiologie et Bio-informatique, Faculté de Sciences et Génie, Université Laval, Québec, QC G1V 0A6, Canada. <sup>2</sup>PROTEO, le regroupement québécois de recherche sur la fonction, l'ingénierie et les applications des protéines, Université Laval, Québec, QC G1V 0A6, Canada. <sup>3</sup>Centre de Recherche en Données Massives (CRDM), Université Laval, Québec, QC G1V 0A6, Canada. <sup>4</sup>Institut de Biologie Intégrative et des Systèmes, Université Laval, Québec, QC G1V 0A6, Canada. <sup>5</sup>Département de Biologie, Faculté de Sciences et Génie, Université Laval, Québec, QC G1V 0A6, Canada. <sup>6</sup>Research Center for Advanced Science and Technology, Synthetic Biology Division, University of Tokyo, Tokyo, 4-6-1 Komaba, Meguro-ku 153-8904, Japan. <sup>7</sup>Department of Biological Sciences, Graduate School of Science, the University of Tokyo, Tokyo, Japan. <sup>8</sup>Institute for Advanced Biosciences, Keio University, Tsuruoka, Japan. ✉email: [yachie@synbiol.rcast.u-tokyo.ac.jp](mailto:yachie@synbiol.rcast.u-tokyo.ac.jp); [christian.landry@bio.ulaval.ca](mailto:christian.landry@bio.ulaval.ca)

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<sup>1,2</sup>. 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 (e.g., complete gene deletion) is applied to all genes<sup>3–5</sup>. 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<sup>6</sup> or by modifying gene expression levels<sup>7–9</sup> 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<sup>10–14</sup>. 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 mutations of interest 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<sup>15–17</sup>. 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)<sup>18</sup> and CRISPR-X<sup>19</sup>, target specific regions of the genome where they induce mutations semi-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<sup>15</sup>, we select gRNAs generating a limited number of predictable edits in yeast essential genes. This allows us to use gRNAs as a readout for the effect of the mutations, similar to commonly used barcode-sequencing approaches to measure fitness effects. Using yeast essential genes as a test case, we identify 708 gRNAs targeting sensitive residues. We then orthogonally validate fitness effects of mutagenesis outcomes using classical genetics approaches and show that they overlap with those predicted to be deleterious. We then use this data to investigate which factors influence base editing efficiency and find multiple gRNAs and target properties that affect mutagenesis and that can be optimized for future experiments in specific genomic spaces.

## 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<sup>20</sup>, 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, and focus on 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<sup>21,22</sup>.

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 study as well as our own work<sup>15,23</sup>. 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 thus 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 positions –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 Fig. 1.

Over 75% of target sequences in this set contained only one or two Cs in the extended activity window (positions –20 to –14), as well as a general enrichment for cytosines in the high activity window (Supplementary Fig. 2A, B). Because the goal of our experiment was to link specific mutations to fitness effects, co-editing of multiple nucleotides using an editor that 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 the “Methods” section). Based on this metric, we found that over 80% of gRNAs fell either in the very low or the low-risk category (Supplementary Fig. 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. 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 Fig. 2D). As Target-AID is known to perform processive editing, a high co-editing risk might also be linked to higher overall editing rate<sup>15</sup>.

**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<sup>24</sup>, 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<sup>15,23</sup>. Furthermore, predicting the activity of gRNAs for base editing remains difficult<sup>25</sup>. 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<sup>23</sup>. 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 Fig. 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 (Fig. 1a).

We included 9 guides from the library isolated from the library quality control process (see the “Methods” section), 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 of 12 gRNAs reaching mutation rates of 50% or higher (Fig. 1b), consistent with previous results<sup>15,23</sup>. Replicates were highly correlated along different measurements with editing rates at the *CAN1* co-editing site being highly consistent (Supplementary Fig. 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.

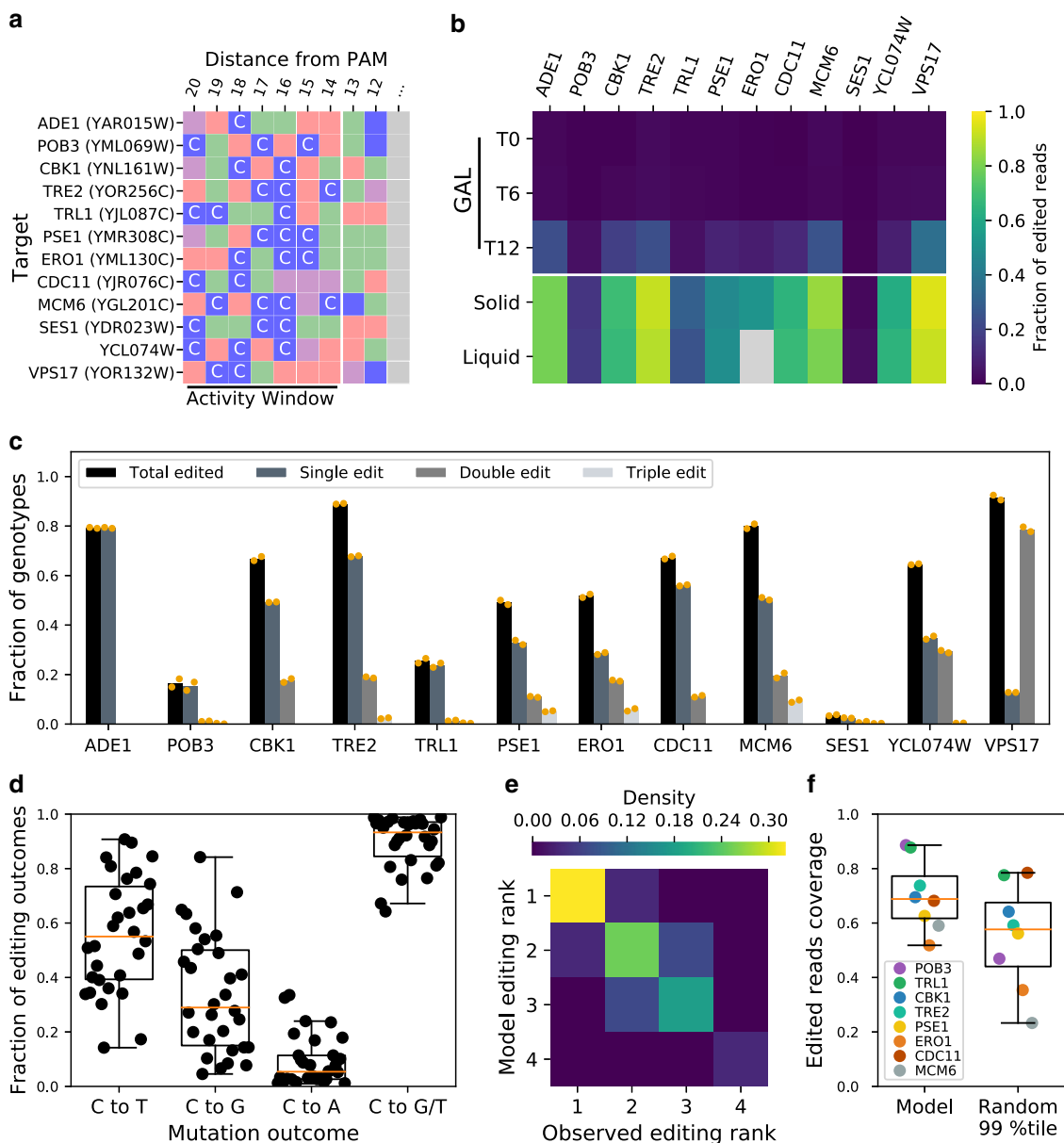
In our editing model, we first predict that single mutants would be the main mutagenesis outcome of the base editing process. We found this to be true for 9 gRNAs out of 10 with more than one cytosine in the Target-AID activity window (Fig. 1c). Second, our model considers C to A editing to be rare and 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 (Fig. 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:  $W = 0$ ,  $p = 1.73 \times 10^{-6}$ , C to G vs C to A:  $W = 41$ ,  $p = 8.19 \times 10^{-5}$ , two-sided Wilcoxon signed-rank test). Including both mutagenesis outcomes as in our model leads to a median coverage of 93%. Our sequencing data also showed a greater prevalence of C to T mutations compared to C to G ( $W = 112$ ,  $p = 0.01$ , two-sided Wilcoxon signed-rank test), but if absolute editing rate is taken into account this difference disappears (Supplementary Fig. 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 (Fig. 1e) which is unlikely to occur by chance ( $p \approx 0.0004$  based on  $1 \times 10^6$  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% (Fig. 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 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<sup>23</sup>. We performed pooled mutagenesis followed by bulk competition (Supplementary Fig. 7) to identify mutations with significant fitness effects (Fig. 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<sup>6,26</sup>.

After applying a stringent filtering threshold based on gRNA read count at the mutagenesis step (see the “Methods” section), 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 Fig. 8) and cluster by experimental step (Supplementary Fig. 9), showing that the approach is reproducible. Using the distribution of abundance variation of gRNAs with synthesis errors (see the “Methods” section) as a null distribution, we identified 708 gRNAs across 605 loci with significant negative effects (GNE) on cell survival or proliferation at a 10% False Discovery Rate (Fig. 3A, Supplementary Fig. 7B and C). GNEs are distributed evenly across the yeast genome (Fig. 3b), suggesting no inherent bias against specific regions. An example of gRNA abundance variation through time for all gRNAs (both GNEs and non-significant gRNAs (NSGs)) targeting *GLN4* is shown in Fig. 3c.

Because our screen specifically targeted essential genes, many gRNAs caused 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<sup>27</sup> in a heterozygous deletion background<sup>28</sup> (Supplementary Fig. 10A). Glycine 267 is part of the “HIGH” motif, characteristic of class I tRNA synthetases, which is involved in ATP binding and catalysis and is highly conserved through evolution<sup>29</sup>. 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 Fig. 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<sup>30</sup>. We found that mutating Gly 267 either to Arg or Ser was enough to cause protein loss of function (Fig. 3d).

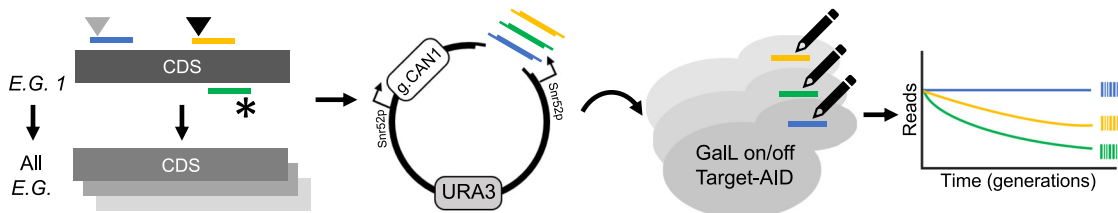
The three other sensitive sites identified in *GLN4* by our screen were also clustered in regions with slow evolutionary rates. We found that a NSG near the z-score threshold targeting residue D291 induced a highly deleterious mutation coupled with a neutral mutation as outcomes (D291E vs D291D, Supplementary Fig. 11). However, we did not observe any discernible growth defect for the other tested GNE outcomes as well as for the outcomes of 6 other NSGs 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. The case of the NSG-induced D291E/D291D pair, where a strong fitness effect is partially obscured by a neutral mutation produced by the



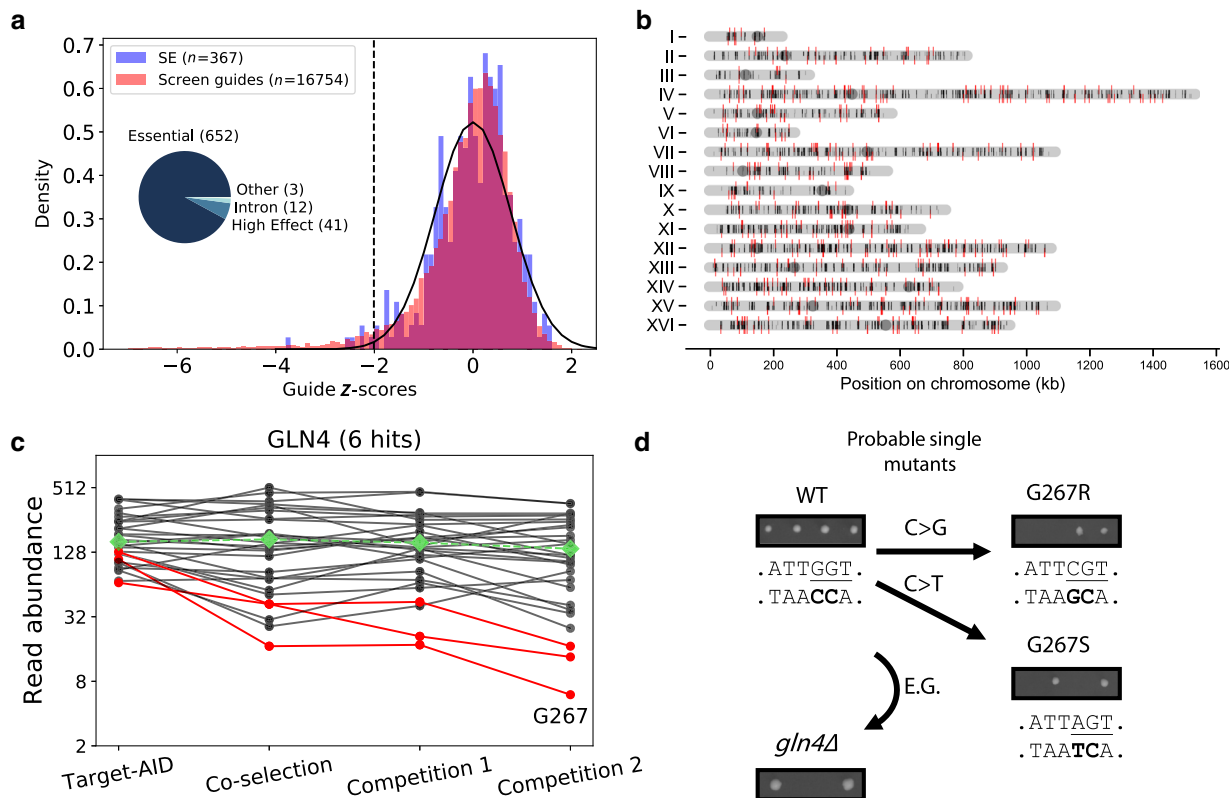
**Fig. 1 A parsimonious model predicts the most probable outcomes of Target-AID mutagenesis.** **a** gRNAs included in the time course base editing experiment had diverse C content profiles in the Target-AID activity window. Nucleotides are color coded: guanines are purple, thymines are red, adenines are green and cytosines are blue. **b** Overall fraction of edited reads for all target sites along time points 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 gray), and as such the solid recovery time point was used instead for the other analysis steps. **c** Fraction of genotypes 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 outcomes are shown in Supplementary Figs. 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: gRNA specific predicted and observed rankings are presented in Supplementary Figs. 5 and 6). 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. Boxplots represent the upper and lower quartiles of the data, with the median shown as a yellow bar. Whiskers extend to 1.5 times the interquartile range (Q3-Q1) at most. Source data are available in the Source Data file.

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 NSGs. We tested this using two recently published resources for variant effect prediction: Envision<sup>2</sup> and Mutfunc<sup>31</sup>. Envision is based on a machine learning approach that leverages



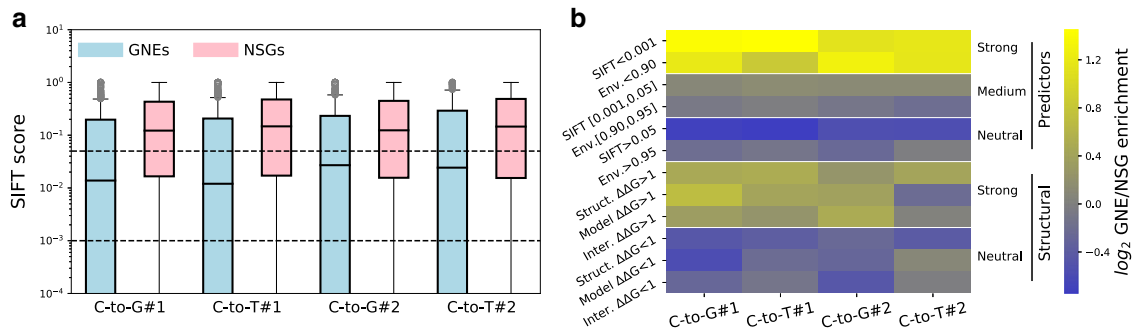
**Fig. 2 A gRNA library for systematic perturbation of essential genes using the Target-AID base editor.** Essential genes (ex.: *E.G.1*) were scanned for sites appropriate for Target-AID mutagenesis. Mutational outcomes include silent (gray triangle), missense (black triangle) mutations, as well as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



**Fig. 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 log<sub>2</sub> fold-change in gRNA abundance between mutagenesis and the end of the bulk competition experiment. Scores were calculated using the distribution of abundance variation of gRNAs with synthesis errors (SE). The fitted normal distribution is shown as a black line, and the 10% FDR threshold as a dotted black line. The distribution of target types in the 708 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** Average decline in gRNA abundance (on a log scale) between time points ( $n = 2$  replicates) after mutagenesis for gRNAs targeting *GLN4* ( $n = 30$  gRNAs), a tRNA synthetase. Median gRNA abundance across the entire library over 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 validates 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. Source data are available in the Source Data file.

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 predicted effect on protein function. Mutfunc aggregates multiple types of information such as residue conservation through the use of SIFT<sup>32</sup> 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 Fig. 12A). The structural modeling information had much lower coverage, covering at best around



**Fig. 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<sup>32</sup> but might be permissive considering the number of mutations tested in our experiment ( $n = 571, 12,718, 457, 8767, 430, 7609, 343, 5847$ ). All GNE vs NSG score comparisons are significant (Welch's  $t$ -test  $p$ -values:  $1.64 \times 10^{-21}$ ,  $5.99 \times 10^{-20}$ ,  $1.62 \times 10^{-12}$ ,  $1.75 \times 10^{-9}$ ). Boxplots represent the upper and lower quartiles of the data, with the median shown as a black bar. Whiskers extend to 1.5 times the interquartile range (Q3-Q1) at most. Outliers are shown in gray. 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 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'. Source data are available in the Source Data file.

12% of the most probable mutations (Supplementary Fig. 12B). As expected, mutations generated by GNEs showed significantly lower SIFT scores and showed enrichment for strong effects predicted by SIFT and Envision (Fig. 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 NSGs. Envision scores across the proteome show a high level of homogeneity, with most mutations having a score between 0.94 and 0.96 (Supplementary Fig. 12C). According to the original Envision study<sup>2</sup>, 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 Fig. 13A).

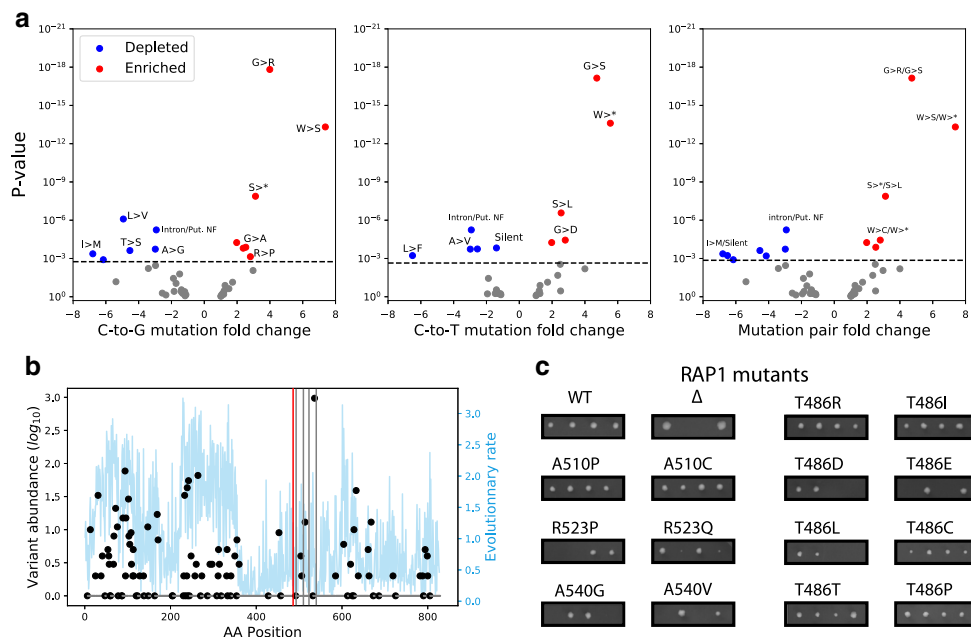
Mutations with destabilizing effects as predicted by structural data also appeared to be enriched in GNEs predicted mutations but low residue coverage limits the strength of this association. This is supported by the raw  $\Delta\Delta G$  value distributions, which show a significant tendency for GNE mutations to be more destabilizing (Welch's  $t$ -test  $p$ -values for GNE vs NSG  $\Delta\Delta G$ : C-to-G #1 0.0001, C-to-T #1 0.0064, C-to-G #2 0.148, C-to-T #2 0.007, Supplementary Fig. 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<sup>33</sup>, resulting in individual destabilizing mutations having a limited effect on fitness. As expected from known experimental data on mutagenesis outcomes<sup>15</sup>, signal was usually stronger for the most probable C to G mutation.

**Sensitive sites provide 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 (Fig. 5a). We found deviations from random expectations in both C-to-G and C-to-T mutation ratios that drove the enrichment of several mutation combinations. 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<sup>34</sup>, making these substitutions highly deleterious. Furthermore, as Glycine residues are often important components of cofactor binding motifs (e.g.,: Phosphates)<sup>35</sup> 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 stop codons: both C-to-G patterns (S to stop: 3.1 fold enrichment,  $p = 3.32 \times 10^{-8}$ , Y to stop: 2.5 fold enrichment,  $p = 0.0001$ ) but only one C-to-T pattern was overrepresented significantly (W to stop, 5.6 fold enrichment,  $p = 2.48 \times 10^{-14}$ ). 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 are enriched: this is not surprising, as the alternative mutations Tryptophan to Serine or Cysteine are also highly disruptive<sup>34</sup>. Changes between similar amino acids, which are expected to be tolerable, were also generally depleted in GNE (ex.: the Alanine to Glycine/Valine pair, 3 fold depletion,  $p = 0.0002$ ). Mutations in intronic sequences and putative non-functional peptides were also underrepresented, as were most patterns leading to silent mutations (Fig. 5a). These results show the power of this approach to discriminate important functional sites from more mutation-tolerant ones across the genome.

Interestingly, genes for which more than one GNE were detected were enriched for molecular function terms linked to cofactor binding (Supplementary Table 1). This suggests that the GNEs might indeed have a tendency to affect protein function through mechanisms other than protein or interaction interface destabilization. These protein properties depend on many residues, making them more robust to single amino acid substitutions, whereas cofactor binding may depend specifically on a handful of residues, making these sites critical for function. Using the Uniprot database<sup>36</sup>, we also examined whether gRNAs that target annotated binding sites or highly conserved motifs are more likely to affect fitness compared to other gRNAs targeting



**Fig. 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 two-sided 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 to correct for multiple testing and is shown as a dotted gray 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 gray line. **c** Tetrad dissections confirm most *RAP1* GNE induced mutations indeed have strong fitness effects, as well as other substitutions targeting these sites. Source data are available in the Source Data file.

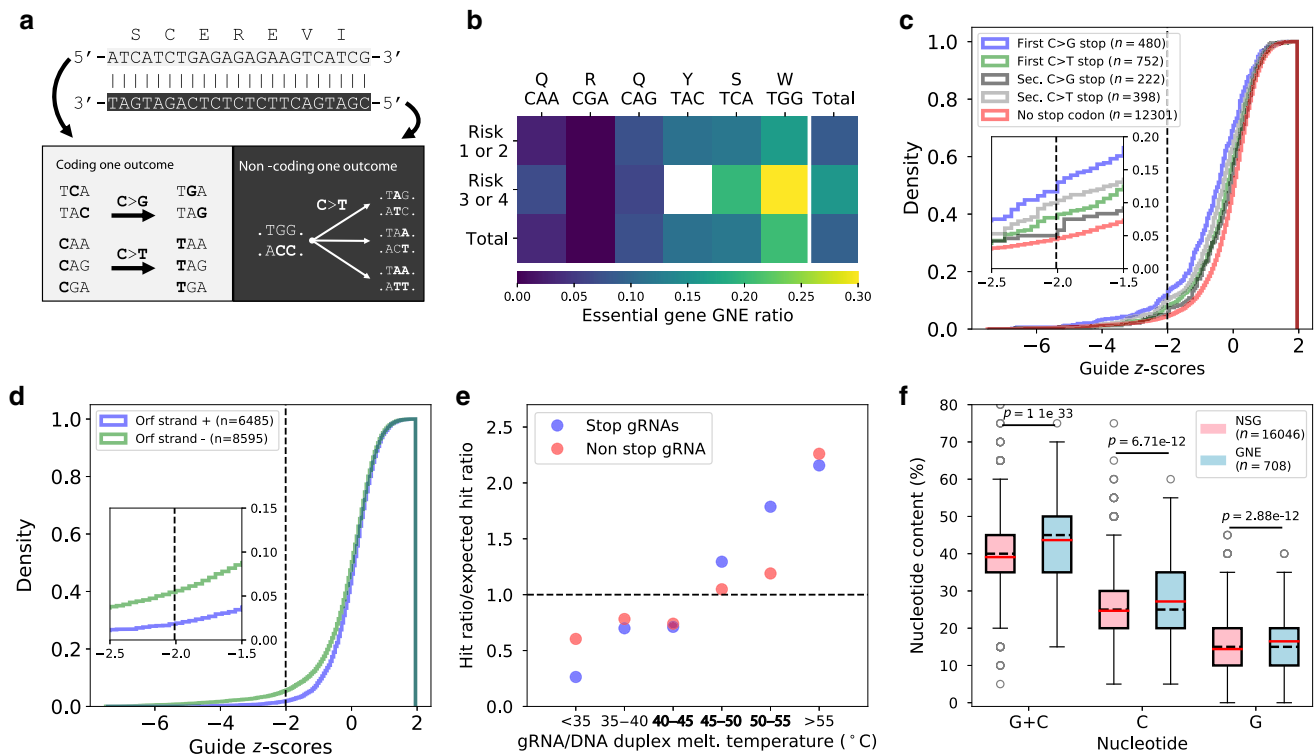
the same set of genes. We found a 4 fold enrichment for GNEs directly affecting these sites (37/188,  $\text{ratio}^{\text{GNE On}} = 0.196$ , 389/5963  $\text{ratio}^{\text{GNE Off}} = 0.048$ , two-sided Fisher’s exact test  $p = 1.45 \times 10^{-12}$ ) or a 2.4 fold enrichment when affecting residues in a two amino acid window around them (17/144,  $\text{ratio}^{\text{GNE near}} = 0.118$ , 389/5963,  $\text{ratio}^{\text{GNE Off}} = 0.048$ , two-sided Fisher’s exact test  $p = 0.00080$ ).

The precise targeting of our method also allowed us to investigate amino acid residues with known functional annotations such as post-translational modifications. We found no significant enrichment for gRNAs mutating directly annotated Post-Translational Modifications (PTMs) ( $\text{ratio}^{\text{GNE PTM}} = 12/708$ ,  $\text{ratio}^{\text{NSG PTM}} = 251/16046$ , Fisher’s exact test  $p = 0.76$ ). Most of these sites were phosphorylation sites (4), metal coordinating residues (3) and ubiquitination sites (2). This is consistent with the hypothesis that many PTM sites may have little functional importance<sup>37</sup> and thus mutations affecting them should not be significantly enriched for strong fitness effects compared to other possible mutations. The same was also observed for gRNAs mutating residues near known PTMs that could disturb recognition sites ( $\text{ratio}^{\text{GNE nearPTM}} = 88/708$ ,  $\text{ratio}^{\text{NSG nearPTM}} = 1763/16046$ , 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<sup>38,39</sup>, but the functional importance of this phosphorylation has not been explored yet. Residue T486 is located in a disordered region in the DNA binding domain<sup>40</sup>,

which is part of the only *RAP1* fragment essential for cell growth<sup>41,42</sup>. Because the available wild-type *RAP1* plasmid (see the “Methods” section) does not complement gene deletion growth phenotype, we used a different strategy for validation that relied on CRISPR-mediated knock-in (see the “Methods” section and Supplementary Fig. 14). We tested the effect of several predicted GNE induced mutations in *RAP1* targeting positions T486, A510, R523 and A540 (Fig. 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 (Fig. 6a). As gRNAs in our library were designed to target the first 75% of the coding



**Fig. 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, where 1 and 2 represent low or very low co-editing risk while 3 or 4 represent moderate to high co-editing risk. **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** GC, C, and G content of NSGs and GNEs. Distribution medians are shown as black dotted lines and means are shown as red lines. *P*-values were calculated using Welch's *t*-tests. Boxplots represent the upper and lower quartiles of the data, and whiskers extend to 1.5 times the interquartile range (Q3-Q1) at most. Outliers are shown in gray.

sequences, successful stop codon generation in this subset of genes should often lead to a lethal loss of function<sup>13,22</sup>.

We found important variation in the ratio of GNE for the different types of SGGs (Fig. 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 a significant GNE enrichment in SGGs which depend on the first C to G mutation to induce stop codon formation (Fig. 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 Fig. 15A). Second, we found a significant enrichment in GNEs for gRNAs targeting the non-coding strand, even after excluding SGGs (Fig. 6d). This effect might be explained by the differences in repair efficiency between strands in yeast<sup>43</sup>. 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 Fig. 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 Fig. 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 (Fig. 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 Fig. 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 Fig. 17). The distribution of GC content in GNEs is shifted compared to NSGs, with a relative increase of 12% (Fig. 6f). Because co-editing tends to increase the mutagenesis efficiency of gRNAs with multiple editable C, this could explain the increased GC content and thus higher melting temperatures we observed for GNEs. However, we found that the relative increase in G content (+14%) in GNEs is higher compared to the relative increase in C content (+10%). This supports that gRNA/DNA binding energy can be an independent factor affecting base editing efficiency.



## Discussion

In previously published methods such as TAM and CRISPR-X<sup>18,19</sup>, 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<sup>15</sup> 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<sup>44,45</sup>. 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<sup>15</sup>. 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<sup>13</sup>. 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 PTC tolerance and position within the coding sequence, something which we also did not observe both for SGGs and non-SGG gRNAs (Supplementary Fig. 17A, B).

We provide key empirical data on gRNA dependant parameters that can be used to optimize base editing efficiency. Based on our results, selecting gRNAs with high binding energy to their genomic targets and favoring those which target the non-coding strand can increase the chance of high editing activity. Importantly, our observations differ from what has been reported for Cas9-based genome editing. High gRNA RNA/DNA duplex binding has instead been associated with lower mutagenesis efficiency<sup>46</sup>. Our data thus confirms the observation that parameters associated with Cas9 editing cannot readily be transferred to base editors<sup>47</sup>. Furthermore, the temperature at which experiments are performed might affect efficiency for certain gRNAs with low gRNA-DNA duplex binding energy and should be considered when designing base editing experiments in different organisms<sup>15</sup>. However, it remains to be confirmed whether the enrichment for certain gRNA properties we observed are specific to Target-AID or will also be transferable to other base editors as this may depend on the enzymatic properties of these proteins. Acquiring large paired gRNA and mutagenesis outcome datasets similar to those available for Cas9 genome editing<sup>24</sup> will allow for more refined models for rational base editing activity prediction.

The field of base editing is rapidly evolving, with new tools being developed constantly. One of the most recent additions to this fast-growing toolkit are engineered Cas9 enzymes with broadened PAM specificities<sup>48</sup>, which have already been shown to be compatible with base editors. More flexible PAM requirements are especially useful for base editing applications, as they increase the number of sites to be edited and also the number of potential gRNAs per site, increasing the chances of choosing optimal properties and thus greater efficiency<sup>25</sup>. Our method allows an experimental scale that bridges saturation mutagenesis methods

and genome-wide knock-out studies, alleviating the current trade-off between mutational diversity and the number of targets genes to generate new biological insights.

## Methods

**Generation of the gRNA library targeting essential genes.** The Target-AID base editor has an activity window between base 15 and 20 in the gRNA sequence starting from the PAM, and the efficiency at these different positions was characterized in Nishida et al.<sup>15</sup>. 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)<sup>49</sup> 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<sup>50</sup>. 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 ( $n = 1156$ )<sup>3,4</sup> 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](http://www-sequence.stanford.edu/group/yeast_deletion_project/Essential_ORFs.txt)). Among non-essential genes, data from Qian et al.<sup>51</sup> 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<sup>23</sup>. 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)<sup>52</sup> and putative non-functional peptides<sup>53</sup> 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 Fig. 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.

**Library construction.** The plasmids, oligonucleotides, and media used in this study are listed in Supplementary Table 2 and Supplementary Data 2 and 3 respectively. The oligo pool was synthesized by Arbor Biosciences (Michigan, USA) and was cloned into the pDYSCO vector using Golden Gate Assembly (New England Biolabs, Massachusetts, USA) with the following reaction parameters: in a 20  $\mu$ l reaction, 2  $\mu$ l NEB GG buffer 10X, 1  $\mu$ l pDYSCO (75 ng  $\mu$ l<sup>-1</sup>), 1  $\mu$ l Oligo pool (2 ng  $\mu$ l<sup>-1</sup>), 1  $\mu$ l NEB GG mix, 15  $\mu$ l PCR grade water. The ligation mix was transformed in *E. coli* strain MC1061 (*[araD139]<sub>B/r</sub> Δ(araA-leu)7697 ΔlacX74 galK16 galE15(Gals) λ-e14-mcrA0 relA1 rpsL150(strR) spoT1 mcrB1 hsdR2*)<sup>54</sup> 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 (Thermo Fisher, Massachusetts, USA), and a normalized master library for yeast transformation was assembled by combining equal quantities of each assembly pool.

**Table 1 Sequence patterns of co-editing risk categories.**

Co-editing risk category	Very Low	Low	Moderate	High
Sequence patterns	NCDDDN	NCDDCN	NCDCNN	NCCDNN
	NDCDDN	NDCDCN		
	NDDCDN	NDDCCN		

*N* any nucleotide, *D* A or T or G.

**Base editing time course and deep sequencing.** Cells were co-transformed with pKNI252 and the pDYSCO 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 Figs. 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  $\mu$ l 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 16,100  $\times$ g and removing the media. Cell pellets were then stored at  $-80^{\circ}\text{C}$  for subsequent DNA extraction. The same method was used to sample ~1 OD at  $T = 6$  h in galactose, ~2 OD at  $T = 12$  h in galactose, and ~3 OD at the end of canavanine co-selection. Plates with selected colonies (edited at the *CANI* 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<sup>55</sup> and quantified by NanoDrop (Thermo Fisher, Massachusetts, USA). For each sample, we aimed to sequence both the target edit site and the *CANI* 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<sup>56</sup>. 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<sup>57</sup>. Reads were then aligned to reference locus sequences using the Needle software from EMBOSS<sup>58</sup>. 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 his3 $\Delta$ 1 leu2 $\Delta$ 0 met15 $\Delta$ 0 ura3 $\Delta$ 0*) cells were first transformed with the pKNI252 (p315-GalL-Target-AID) plasmid using a standard lithium acetate method<sup>59</sup>. Briefly, 50 ml of yeast cells in exponential growth (0.6–0.7 OD) were pelleted (500  $\times$ g, 5 min), washed with 5 ml sterile water, then washed with 5 ml SORB solution (100 mM lithium acetate, 10 mM Tris pH 8.0, 1 mM EDTA, 1 M sorbitol). The cells were then resuspended in 360  $\mu$ l SORB and 40  $\mu$ l carrier DNA (salmon sperm DNA 10 mg ml<sup>-1</sup>). Cells were then aliquoted and stored at  $-80^{\circ}\text{C}$  until use. After thawing, each 20  $\mu$ l aliquot was incubated with ~1  $\mu$ g plasmid DNA and 100  $\mu$ l Plate mixture (PEG 3350 40%, Lithium acetate 10 mM, Tris-Cl pH 7.5 10 mM, EDTA 1 mM) at room temperature for 30 min. DMSO (15  $\mu$ l) was then added and cells are incubated at 42  $^{\circ}\text{C}$  for 20 min. The cells were then pelleted (400  $\times$ g, 3 min) and the supernatant removed before being resuspended in 100  $\mu$ l YPD media for a 4 h outgrowth incubation at 30  $^{\circ}\text{C}$  without shaking. Transformants were selected by plating cells on SC-L. After 48 h of growth, multiple colonies were used to inoculate a starter liquid culture for competent cells preparation using the same protocol: 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  $\mu$ l of competent cells and 5  $\mu$ l of plasmid library (240 ng  $\mu$ l<sup>-1</sup>) after the outgrowth stage and plating 100  $\mu$ l aliquots on SC-UL: cells were then allowed to grow at 30  $^{\circ}\text{C}$  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  $\times$  10<sup>6</sup> 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<sup>23</sup> and is shown in Supplementary Fig. 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  $\times$  10<sup>8</sup> 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 h before

1  $\times$  10<sup>9</sup> cells were pelleted and used to inoculate a 100 ml SC-UL + 5% glycerol culture. After 24 h, 5  $\times$  10<sup>8</sup> 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 pKNI252) was induced for 12 h before 1  $\times$  10<sup>8</sup> cells were pelleted and used to inoculate a canavanine (50  $\mu$ g ml<sup>-1</sup>) co-selection culture in SC-ULR. After 16 h of incubation, 5  $\times$  10<sup>7</sup> cells of each culture were used to inoculate 100 ml SC-UR, which was grown for 12 h before 5  $\times$  10<sup>7</sup> cells were used to inoculate a final 100 ml SC-UR culture which was grown for another 12 h. Cell pellets were washed with sterile water between each step, and all incubation occurred at 30  $^{\circ}\text{C}$  with agitation. ~2  $\times$  10<sup>7</sup> 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 h at room temperature, 1 min at  $-80^{\circ}\text{C}$ , and then incubated for another 15 min at room temperature before the lysis step. Plasmid DNA was eluted in 70  $\mu$ l of E5 buffer (10 mM Tris-HCl, pH 8.5) and stored at  $-20^{\circ}\text{C}$  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 pDYSCO plasmid pool and the second to add sample barcodes as well as the Illumina p5 and p7 sequences<sup>60</sup>. Oligonucleotides for library preparation are shown in the first part of the oligonucleotide table. Reaction conditions for the first PCR were as follows (25  $\mu$ l reaction): 5  $\mu$ l Phusion HF buffer (NEB) 5X, 0.5  $\mu$ l dNTPs 10 mM, 1.25  $\mu$ l pDYSCO\_gRNA\_for 10  $\mu$ M, 1.25  $\mu$ l pDYSCO-O\_gRNA\_rev 10  $\mu$ M, 0.5  $\mu$ l Phusion polymerase, 5  $\mu$ l Template DNA (< 1 ng/ $\mu$ l), 11.7  $\mu$ l PCR grade water. The following thermocycler protocol was used: 98  $^{\circ}\text{C}$  for 30 s, then 16 cycles of 98  $^{\circ}\text{C}$  for 10 s, 58  $^{\circ}\text{C}$  for 15 s, 72  $^{\circ}\text{C}$  for 5 s and then 72  $^{\circ}\text{C}$  for 5 s. The resulting product was 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). The purified products were used as the template for the second PCR reaction, with the following conditions (20  $\mu$ l reaction): 10  $\mu$ l Phusion Mastermix-HF (NEB), 3.75  $\mu$ l P5-barcode-X oligo 1.333  $\mu$ M, 3.75  $\mu$ l P7-barcode-Y oligo 1.333  $\mu$ M, 2.5  $\mu$ l Template DNA (~1 ng/ $\mu$ l). The following thermocycler protocol was then used: 98  $^{\circ}\text{C}$  for 30 s, then 15 cycles of 98  $^{\circ}\text{C}$  for 10 s, 60  $^{\circ}\text{C}$  for 10 s, 72  $^{\circ}\text{C}$  for 1 min and then 72  $^{\circ}\text{C}$  for 5 min. 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 data are available on github ([https://github.com/Landrylab/Despres\\_et\\_al\\_2020](https://github.com/Landrylab/Despres_et_al_2020)), and packages and software used are presented in Supplementary Table 3. Raw sequencing files have been deposited on the NCBI SRA, accession number PRJNA552472. Briefly, reads were separated into three sub-sequences for alignment: the P5 barcode, the gRNA, and the P7 barcode. Each of these was aligned using Bowtie<sup>50</sup> to an artificial reference genome containing either the barcodes or gRNA sequences flanked by the common amplicon sequences. The gRNA sequences were 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:

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

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 ( $N_{\text{readsformismatch}} > 1000$ ) SE gRNAs were kept to serve as negative controls when measuring fitness effects, resulting in a set of 1032 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<sup>61</sup>. 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:

$$\Delta \log_2_{\text{gRNA}} = \log_2 \left( \frac{N_{\text{readsgRNA}_{t_1}}}{N_{\text{readst}_1}} \right) - \log_2 \left( \frac{N_{\text{readsgRNA}_{t_0}}}{N_{\text{readst}_0}} \right) \quad (1)$$

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 less prevalent in the SE gRNAs. We then averaged z-scores between replicates and fit a normal distribution to the distribution of SE gRNA z-scores. We used then used this distribution to estimate the False Positive Rate as a function of gRNA z-score. This, in turn, allowed us to calculate a False Discovery Rate based on the number of hits at a set FPR threshold (Supplementary Fig. 9B and C). We set the z-score threshold so that the FDR = 10%.

**Complementation assays.** Experiments were performed in heterozygous deletion mutants from the YKO project heterozygous deletion strain set (Dharmacon, Colorado, USA). For each gene, a single colony streaked from the glycerol stock was used to prepare competent cells using the previously described lithium acetate protocol<sup>59</sup>. To generate mutant alleles of the genes of interest, we performed site-directed mutagenesis on the appropriate MoBY collection plasmid<sup>27</sup>. These centromeric plasmids encode the yeast gene of interest under the control of their native promoters and terminators. Mutagenesis reactions were performed with the following reaction setup for a 25  $\mu\text{l}$  reaction: 5  $\mu\text{l}$  Kapa HiFi buffer (Kapa Biosciences) 5X, 0.75  $\mu\text{l}$  dNTPs 10  $\mu\text{M}$ , 0.75  $\mu\text{l}$  mutation\_for 10  $\mu\text{M}$  (see Supplementary Data 2), 0.75  $\mu\text{l}$  mutation\_rev 10  $\mu\text{M}$  (see Supplementary Table 7), 0.5  $\mu\text{l}$  Kapa Hot-start polymerase, 0.75  $\mu\text{l}$  Template plasmid DNA (15  $\text{ng } \mu\text{l}^{-1}$ ), 16.5  $\mu\text{l}$  PCR grade water. The following thermocycler protocol was then used: 95  $^{\circ}\text{C}$  for 5 min, then 20 cycles of 98  $^{\circ}\text{C}$  for 20 s, 60  $^{\circ}\text{C}$  for 15 s, 72  $^{\circ}\text{C}$  for 12 min and then 72  $^{\circ}\text{C}$  for 18 min. After amplification, the mutagenesis product was digested with DpnI for 2 h at 37  $^{\circ}\text{C}$  and 5  $\mu\text{l}$  was transformed in *E. coli* strain BW23474 (F<sup>-</sup>,  $\Delta(\text{argF-lac})169$ ,  $\Delta\text{uidA}::\text{pir-116}$ , *recA1*, *rpoS396(Am)*, *endA9(del-ins)::FRT*, *rph-1*, *hsdR514*, *rob-1*, *creC510*)<sup>62</sup>. Transformants were plated on 2YT + Kan [50  $\mu\text{g } \text{ml}^{-1}$ ] + Chlo [12.5  $\mu\text{g } \text{ml}^{-1}$ ] and grown at 37  $^{\circ}\text{C}$  overnight. Plasmid DNA was then isolated from clones and sent for Sanger sequencing (CHUL sequencing platform, Université Laval, Québec City, Canada) to confirm mutagenesis success.

Competent cells of target genes were transformed with the appropriate mutant plasmids as well as the original plasmid bearing the wild-type gene and the empty vector<sup>63</sup>, and transformants were selected by plating on SC-U (MSG). Multiple independent colonies per transformation were then put on sporulation media until sporulation could be confirmed by microscopy. For tetrad dissection, cells were resuspended in 100  $\mu\text{l}$  20T zymolyase (200  $\mu\text{g } \text{ml}^{-1}$  dilution in water) and incubated for 20 min at room temperature. Cells were then centrifuged and resuspended in 50  $\mu\text{l}$  1 M sorbitol before being streaked on a level YPD plate. All dissections were performed using a Singer SporePlay microscope (Singer Instruments, UK). Plate pictures were taken after 5 days incubation at room temperature except for the *RAP1* plasmid complementation test for which the picture was taken after 3 days. Pictures are shown in Source Image File 1 in the Source Data.

**Strain construction for confirmations in *RAP1*.** Because the MoBY collection plasmid for *RAP1* cannot fully complement the gene deletion (Source image file 1 in the Source Data.), we instead performed confirmations by engineering mutations in a diploid strain to create heterozygous mutants. *RAP1* was first tagged with a modified version of fragment DHFR F[1,2] (the first half) of the mDHFR enzyme<sup>64</sup>. The mDHFR[1,2]-FLAG cassette was amplified using gene-specific primers and previously described reaction parameters<sup>64</sup>. Cells were transformed with the cassette using the previously described transformation protocol and were plated on YPD + Nourseothricine (YPD + Nat in Media table). Positive clones

were identified by colony PCR and successful fragment fusion was confirmed by Sanger sequencing (CHUL sequencing platform). We then mated the confirmed clones with strain Y8205 (*Mata can1::STE2pr-his5 lyp1::STE3prLEU2  $\Delta$ ura3  $\Delta$ his3  $\Delta$ leu2*, kindly gifted by Charlie Boone) by inoculating a 4 ml 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<sup>65</sup>.

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 Fig. 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<sup>66</sup>. 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, Université 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 5 days at room temperature (Source Image File 2 in the Source Data.).

**Evolutionary rate measurements and variant abundance.** Evolutionary rates were calculated using the Rate4site software<sup>67</sup> using multiple sequence alignments and phylogenies from PhylomeDB V4<sup>68</sup> 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<sup>69</sup> 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<sup>70</sup>. 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<sup>71</sup>, which uses values taken from Sugimoto et al.<sup>72</sup>. Correlation between gRNA/DNA duplex melting temperatures was assessed using Spearman's rank correlation.

**Variant effect prediction analysis and GO enrichment.** All prediction data except the Envision scores were extracted from the aggregated data of the Mutfunc database<sup>31</sup>. Precomputed values were downloaded directly from the FTP server ([http://ftp.ebi.ac.uk/pub/databases/mutfunc/mutfunc\\_v1/yeast/](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  $\Delta\Delta G$  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<sup>2</sup> were downloaded directly from the database website ([https://envision.gs.washington.edu/shiny/envision\\_new/](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 Fig. 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<sub>All</sub> = 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<sup>73</sup>. 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.

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

### Data availability

All raw sequencing data has been deposited on the NCBI BioProject as accession number PRJNA552472. The gRNA screen scores, predicted mutation outcomes, mutation effect predictors scores, as well as other relevant annotations are provided in Supplementary Data 1. Source image files for the tetrad dissections are presented as Source Image 1 and 2. The data underlying each figure is detailed in the source data file. Data on yeast gene essentiality was obtained from [http://www-sequence.stanford.edu/group/yeast\\_deletion\\_project/](http://www-sequence.stanford.edu/group/yeast_deletion_project/). The mutfunc database is available at [http://ftp.ebi.ac.uk/pub/databases/mutfunc/mutfunc\\_v1/yeast/](http://ftp.ebi.ac.uk/pub/databases/mutfunc/mutfunc_v1/yeast/). The Envision database is available at [https://envision.gs.washington.edu/shiny/envision\\_new/](https://envision.gs.washington.edu/shiny/envision_new/). The Uniprot database can be found at <https://www.uniprot.org/>. The Saccharomyces database can be accessed at <https://www.yeastgenome.org/>. The PANTHER gene ontology tool can be accessed at <http://pantherdb.org/>. Phylome DB V4 can be accessed at <http://phylomedb.org/>. The data from the 1002 yeast genome project can be accessed at <http://1002genomes.u-strasbg.fr/files/allReferenceGenesWithSNPsAndIndelsInferred.tar.gz>. All other relevant data are available from the authors upon reasonable request.

### Code availability

The custom Python scripts used to analyze the are available on github ([https://github.com/Landrylab/Despres\\_et\\_al\\_2020](https://github.com/Landrylab/Despres_et_al_2020)), and packages and software used are presented in Supplementary Table 3. The base editing deep sequencing data was obtained using a MiSeq sequencer that uses proprietary Illumina software to generate fastq files. The high-throughput sequencing data for the genome-wide screen was collected using a HiSeq 2500 sequencer that uses proprietary Illumina software to generate fastq files. The following programs were used during data analysis: PANDA-Seq v2.11, Needle (EMBOSS v6.6.0.0), Bowtie v1.2.1.1, Rate4site 3.0.0, and FASTX v36.3.8. The following packages for Python 2.7 were used: pandas 0.23.4, matplotlib v2.2.3, numpy v1.15.4, scipy v1.1.0, seaborn v0.9.0, tqdm v4.32, and Biopython v1.71. For Python 3.6, the following packages were used: pandas v0.24.2, numpy v1.16.2, matplotlib v3.0.3, scipy v1.2.1 and seaborn v0.9.0.

Received: 18 July 2019; Accepted: 27 March 2020;

Published online: 20 April 2020

### References

- Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
- 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.* **6**, 116–124.e3 (2018).
- Winzler, E. A. et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
- Giaever, G. et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
- The C. elegans Deletion Mutant Consortium. Large-scale screening for targeted knockouts in the *Caenorhabditis elegans* Genome. *G3 (Bethesda)* **2**, 1415–1425 (2012).
- Shalem, O. et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
- Qi, L. S. et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).
- Sander, J. D. & Joung, J. K. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32**, 347–355 (2014).
- 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).
- Sharon, E. et al. Functional genetic variants revealed by massively parallel precise genome editing. *Cell* **175**, 544–557.e16 (2018).
- Bao, Z. et al. Genome-scale engineering of *Saccharomyces cerevisiae* with single-nucleotide precision. *Nat. Biotechnol.* **36**, 505–508 (2018).
- Roy, K. R. et al. Multiplexed precision genome editing with trackable genomic barcodes in yeast. *Nat. Biotechnol.* **36**, 512–520 (2018).
- Sadhu, M. J. et al. Highly parallel genome variant engineering with CRISPR-Cas9. *Nat. Genet.* **50**, 510–514 (2018).
- Guo, X. et al. High-throughput creation and functional profiling of DNA sequence variant libraries using CRISPR-Cas9 in yeast. *Nat. Biotechnol.* **36**, 540–546 (2018).
- Nishida, K. et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **353**, 553–563 (2016).
- 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).
- 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).
- Ma, Y. et al. Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. *Nat. Methods* **13**, 1029–1035 (2016).
- Hess, G. T. et al. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods* **13**, 1036–1042 (2016).
- Giaever, G. et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
- Doench, J. G. et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
- Michel, A. H. et al. Functional mapping of yeast genomes by saturated transposition. *Elife* **6**, e23570 (2017).
- 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)* **8**, 3163–3171 (2018).
- Allen, F. et al. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.* <https://doi.org/10.1038/nbt.4317> (2019).
- 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).
- Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods* **11**, 783–784 (2014).
- 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).
- Giaever, G. et al. Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat. Genet.* **21**, 278–283 (1999).
- 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).
- 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.* <https://doi.org/10.1093/nar/23.23.4793> (1995).
- Wagih, O. et al. A resource of variant effect predictions of single nucleotide variants in model organisms. *Mol. Syst. Biol.* **14**, e8430 (2018).
- Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
- 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).
- Sneath, P. H. Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* **12**, 157–195 (1966).
- Copley, R. R. & Barton, G. J. A Structural Analysis of Phosphate and Sulphate Binding Sites in Proteins. *J. Mol. Biol.* **242**, 321–329 (1994).
- Bateman, A. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gky1049> (2019).
- Landry, C. R., Levy, E. D. & Michnick, S. W. Weak functional constraints on phosphoproteomes. *Trends Genet.* **25**, 193–197 (2009).
- Albuquerque, C. P. et al. A multidimensional chromatography technology for in-depth phosphoproteome analysis. *Mol. Cell. Proteom.* **7**, 1389–1396 (2008).
- Holt, L. J. et al. Global analysis of Cdk1 substrate phosphorylation sites provides insights into evolution. *Science* **325**, 1682–1686 (2009).
- Konig, P., Giraldo, R., Chapman, L. & Rhodes, D. The crystal structure of the DNA-binding domain of yeast RAP1 in complex with telomeric DNA. *Cell* **85**, 125–136 (1996).
- Graham, I. R., Haw, R. A., Spink, K. G., Halden, K. A. & Chambers, A. In vivo analysis of functional regions within yeast Rap1p. *Mol. Cell. Biol.* **19**, 7481–7490 (1999).
- Wu, A. C. K. et al. Repression of divergent noncoding transcription by a sequence-specific transcription factor. *Mol. Cell* **72**, 942–954.e7 (2018).
- Reis, A. M. C. et al. Targeted detection of in vivo endogenous DNA base damage reveals preferential base excision repair in the transcribed strand. *Nucleic Acids Res.* **40**, 206–219 (2012).

44. Jin, S. et al. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. *Science* **364**, 292–295 (2019).
45. Zuo, E. et al. Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. *Science* <https://doi.org/10.1126/SCIENCE.AAV9973> (2019).
46. Wong, N., Liu, W. & Wang, X. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biol.* **16**, 218 (2015).
47. Kim, D. et al. Genome-wide target specificities of CRISPR RNA-guided programmable deaminases. *Nat. Biotechnol.* <https://doi.org/10.1038/nbt.3852> (2017).
48. Nishimasu, H. et al. Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science* **361**, 1259–1262 (2018).
49. Dicarlo, J. E. et al. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Res.* **41**, 4336–4343 (2013).
50. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
51. Qian, W., Ma, D., Xiao, C., Wang, Z. & Zhang, J. The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast. *Cell Rep.* **2**, 1399–1410 (2012).
52. Grate, L. & Ares, M. Searching yeast intron data at Ares lab web site. *Methods Enzymol.* **350**, 380–392 (2002).
53. Smith, J. E. et al. Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. *Cell Rep.* **7**, 1858–1866 (2014).
54. Casadaban, M. J. & Cohen, S. N. Analysis of gene control signals by DNA fusion and cloning in *Escherichia coli*. *J. Mol. Biol.* **138**, 179–207 (1980).
55. Amberg, D. C., Burke, D. J. & Strathern, J. N. Methods in yeast genetics: A Cold Spring Harbor Laboratory Course Manual, 2005 Edition. A Cold Spring Harbor Laboratory Course Manual (2005).
56. Yachie, N. et al. Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol. Syst. Biol.* <https://doi.org/10.15252/msb.20156660> (2016).
57. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* <https://doi.org/10.1186/1471-2105-13-31> (2012).
58. Rice, P., Longden, L. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* [https://doi.org/10.1016/S0168-9525\(00\)02024-2](https://doi.org/10.1016/S0168-9525(00)02024-2) (2000).
59. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).
60. Yachie, N. et al. Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol. Syst. Biol.* **12**, 863 (2016).
61. Robinson, D. G., Chen, W., Storey, J. D. & Gresham, D. Design and analysis of Bar-seq experiments. *G3 (Bethesda)* **4**, 11–18 (2014).
62. Haldimann, A. et al. Altered recognition mutants of the response regulator PhoB: a new genetic strategy for studying protein-protein interactions. *Proc. Natl Acad. Sci. USA* **93**, 14361–14366 (1996).
63. Zhao, L. et al. A genome-wide imaging-based screening to identify genes involved in synphilin-1 inclusion formation in *Saccharomyces cerevisiae*. *Sci. Rep.* **6**, 30134 (2016).
64. Tarassov, K. et al. An in vivo map of the yeast protein interactome. *Science* **320**, 1465–1470 (2008).
65. Huxley, C., Green, E. D. & Dunham, I. Rapid assessment of *S. cerevisiae* mating type by PCR. *Trends Genet.* **6**, 236 (1990).
66. Ryan, O. W., Poddar, S. & Cate, J. H. D. Crispr-cas9 genome engineering in *Saccharomyces cerevisiae* cells. *Cold Spring Harb. Protoc.* **2016**, 525–533 (2016).
67. Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. Comparison of site-specific rate-inference methods for protein sequences: empirical bayesian methods are superior. *Mol. Biol. Evol.* **21**, 1781–1791 (2004).
68. Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M. & Gabaldón, T. PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.* **42**, D897–D902 (2014).
69. Pearson, W. R., Wood, T., Zhang, Z. & Miller, W. Comparison of DNA sequences with protein sequences. *Genomics* **46**, 24–36 (1997).
70. Cherry, J. M. et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkr1029> (2012).
71. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btp163> (2009).
72. Sugimoto, N. et al. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry* **34**, 11211–11216 (1995).
73. Mi, H. et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat. Protoc.* **14**, 703–721 (2019).

## Acknowledgements

This work was supported by the Canadian Institutes of Health Research Foundation grant 387697 as well as project grants 364920, 384483 to C.R.L. and by 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 paper, as well as Maria Isabel Acosta Lopez for assistance during the strain construction process and Ugo Dionne for help with plasmid construction.

## Author contributions

P.C.D., A.K.D., N.Y., and C.R.L. designed research. P.C.D. and A.K.D. performed experiments. P.C.D. and M.S. generated NGS sequencing data. All data analysis was performed by P.C.D. with input from C.R.L. P.C.D. and C.R.L. wrote the paper with input from all authors.

## Competing interests

The authors declare no competing interests.

## Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41467-020-15796-7>.

Correspondence and requests for materials should be addressed to N.Y. or C.R.L.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020