



HHS Public Access

Author manuscript

Nat Biotechnol. Author manuscript; available in PMC 2018 November 07.

Published in final edited form as:

Nat Biotechnol. 2018 July ; 36(6): 512–520. doi:10.1038/nbt.4137.

Multiplexed precision genome editing with trackable genomic barcodes in yeast

Kevin R. Roy^{1,2,3,4,†}, Justin D. Smith^{1,4,†}, Sibylle C. Vonesch^{5,†}, Gen Lin⁵, Chelsea Szu Tu⁵, Alex R. Lederer⁵, Angela Chu^{1,6}, Sundari Suresh^{1,6}, Michelle Nguyen^{1,4}, Joe Horecka^{1,6}, Ashutosh Tripathi⁷, Wallace T. Burnett^{1,4}, Maddison A. Morgan^{1,4}, Julia Schulz^{1,4}, Kevin M. Orsley^{1,4}, Wu Wei^{1,4}, Raeka S. Aiyar¹, Ronald W. Davis^{1,4,6}, Vytas A. Bankaitis^{7,8,9}, James E. Haber¹⁰, Marc L. Salit^{2,3}, Robert P. St Onge^{1,6,*}, and Lars M. Steinmetz^{1,3,4,5,*}

¹Stanford Genome Technology Center, Stanford University, Palo Alto, California, USA

²Genome-Scale Measurements Group, Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland, USA

³Joint Initiative for Metrology in Biology, Stanford, California, USA

⁴Department of Genetics, Stanford University School of Medicine, Stanford, California, USA

⁵European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany

⁶Department of Biochemistry, Stanford University School of Medicine, Stanford, California, USA

⁷Department of Molecular and Cellular Medicine, Texas A&M Health Science Center, College Station, Texas, USA

⁸Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas, USA

⁹Department of Chemistry, Texas A&M University, College Station, Texas, USA

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: http://www.nature.com/authors/editorial_policies/license.html#terms

*correspondence to Robert P. St. Onge (bstonge@stanford.edu) or Lars M. Steinmetz (larsms@stanford.edu).

†These authors contributed equally to this work

Author Contributions

K.R.R., J.D.S., S.C.V., R.P.S. and L.M.S conceived and designed the study, and wrote and edited the paper. K.R.R., J.D.S., S.C.V., and R.P.S. performed experiments and analyzed data. K.R.R., S.C.V., G.L., and A.R.L. analyzed NGS data; C.S.T., A.C., S.S., M.N., J.H., W.T.B., M.A.M., J.S., and K.M.O. performed experiments. A.T. and V.A.B. performed computational structural analysis on Sec14p-NPPM; W.W. performed variant calling for the different yeast strains. J.E.H. suggested adapting the LexA-Fkh1p system to the guide-donor plasmid. R.S.A., R.W.D., and M.L.S. advised the study. R.P.S. and L.M.S were responsible for the coordination of the study. All authors read, corrected and approved the final manuscript.

Competing interests

K.R.R., J.D.S., J.E.H., R.P.S. and L.M.S have filed a provisional application (US 62/559,493) with the US Patent and Trademark Office on this work.

Accessions

Protein Data Bank: 1AUA (open conformer of Sec14p), 3B7N (closed conformer of Sfh1p)

Code Availability

All code used to design guide-donor oligo libraries and analyze sequencing data is available here: <https://github.com/k-roy/MAGESTIC>

Data Availability

All sequencing data associated with this study (Fig. 2a, Fig. 4c, Fig. 5b,c,d) is available from the European Nucleotide Archive (ENA) under primary accession number **PRJEB23616**.

¹⁰Rosenstiel Basic Medical Sciences Research Center and Department of Biology, Brandeis University, Waltham, Massachusetts, USA

Abstract

Our understanding of how genotype controls phenotype is limited by the scale at which we can precisely alter the genome and assess phenotypic consequences of each perturbation. Here we describe a CRISPR/Cas9-based method for multiplexed accurate genome editing with short, trackable, integrated cellular barcodes (MAGESTIC) in *S. cerevisiae*. MAGESTIC uses array-synthesized guide-donor oligos for plasmid-based high-throughput editing and features genomic barcode integration to prevent plasmid barcode loss and to enable robust phenotyping. We demonstrate that editing efficiency can be increased >5-fold by recruiting donor DNA to the site of breaks using the LexA-Fkh1p fusion protein. We performed saturation editing of the essential gene *SEC14* and identified amino acids critical for chemical inhibition of lipid signaling. We also constructed thousands of natural genetic variants, characterized guide mismatch tolerance at the genome-scale, and ascertained that cryptic Pol III termination elements substantially reduce guide efficacy. MAGESTIC will be broadly useful to uncover the genetic basis of phenotypes in yeast.

Predicting the functional consequences of genetic variation is one of the fundamental challenges in understanding phenotypic diversity, engineering desirable traits for biotechnology, and enabling precision medicine. Although CRISPR screens have been used extensively to disrupt function through the introduction of non-homologous end-joining (NHEJ)-mediated small insertions/deletions (indels) and premature termination codons (PTCs) in open reading frames (ORFs), few methods have been developed to introduce specific amino acid and nucleotide variants at the genome scale.

High-throughput approaches for genome editing have been described in prokaryotes¹ and more recently in yeast^{2,3}, but these studies have not explored natural genetic variation. Here we describe a CRISPR/Cas9-based method in *S. cerevisiae* for multiplexed genome editing with array-synthesized guide RNA/donor DNA (guide-donor) oligonucleotides that overcomes major shortcomings in currently employed approaches. First, we introduce stable, genome-integrated barcodes instead of plasmid barcodes, thereby enabling marker-free variant tracking and one-to-one correspondence of barcode counts to strain abundance. Second, we demonstrate >5-fold increase in precision editing efficiency by active recruitment of donor DNA to Cas9-induced double strand breaks. This improvement enabled saturating a region of the *SEC14* gene with all possible amino acid changes to identify residues modulating sensitivity to the NPPM class of Sec14p-like phosphatidylinositol transfer protein inhibitors, attractive drug targets in pathogenic fungi. Finally, we use MAGESTIC to introduce thousands of SNPs and small indels, representing natural variants from the vineyard isolate RM11, into the laboratory strain S288c. We demonstrate the ability to make single-nucleotide variants without requiring PAM mutations, reveal distinct mismatch tolerance between the 19th and 20th positions from the PAM, and ascertain that the presence of cryptic Pol III termination signals in the form of imperfect T-homopolymer stretches is a key factor predicting guide efficiency.

The MAGESTIC workflow

MAGESTIC utilizes pools of array-synthesized oligos encoding a guide RNA, a Type IIS restriction site, and a donor DNA to introduce the designed variant by homologous recombination (HR; Fig. 1a). The guide-donor pairs enable multiplexed engineering of specific genetic variants at desired locations throughout the genome and quantification of variant abundance by sequencing. Synthesis errors in the guide sequence can prevent target recognition and cleavage, and errors in the donor DNA can lead to incorporation of the wrong variant. To enable accurate phenotyping free from confounding sequencing errors, we tagged each guide-donor pair with a short, unique 31-mer barcode during subpool amplification. Paired-end sequencing assigns each barcode to its corresponding guide-donor sequence and enables full-length sequence verification (Fig. 1a). Multiple distinct barcodes mapping to the same guide-donor combination offer the further advantage of internal replicates for a given edit, and can be leveraged as single-cell barcodes⁴. The structural component of the Cas9 guide as well as bacterial- and yeast-specific selectable markers are added via the Type IIS-site in a second cloning step (Fig. 1b). Selecting for these markers in step 2 cloning removes uncut step 1 products and ensures a high-quality library, which is then transformed into a population of yeast cells harboring Cas9⁵.

To enable one-to-one barcode-to-cell correspondence and to eliminate the need for plasmid maintenance, the guide-donor cassette is linearized and integrated into the genome using a dedicated guide (guide X) targeting both the plasmid and a chromosomal barcode locus; insertion is mediated by identical homologies flanking the guide-donor on the plasmid and barcode locus (Fig. 1b). The abundance of each variant is assessed after competitive growth in different conditions by next-generation sequencing (NGS) of the 31-mer barcodes, enabling high-throughput profiling of variant function (Fig. 1c). To test whether library diversity and uniformity are maintained throughout each step of the pipeline, we cloned a guide-donor library harboring 10^5 members, achieving 2×10^6 distinct barcodes (~20-fold library coverage) in the first step of cloning. Of the 10^5 designed guide-donors, we identified 99% and 94% after the first and second cloning steps, respectively, and 89% after yeast transformation without substantial increase in bias (Supplementary Fig. 1).

Simultaneous editing, barcoding, and plasmid destruction

As a proof of principle, we designed a guide-donor plasmid to introduce a PTC into the *ADE2* ORF (*YOR128C*). Disruption of *ADE2* results in accumulation of red pigment, enabling direct visual identification of edited colonies⁶. First we tested three different Pol III promoters (*RPR1*, *SNR52*, and tRNA-Tyr(*SUP4*)-HDV) to drive expression of guide X, and found similar kinetics of barcode integration upon Cas9 induction (Supplementary Fig. 2). To assess editing kinetics with MAGESTIC, we quantified the fraction of NHEJ indel and HR donor repair events in the population by NGS of the *ADE2* locus throughout 15 generations of editing (Fig. 2a). Over 9 generations, perfect donor repair events approached 70% with the remaining 30% constituting NHEJ indels (Fig. 2a). Precision donor editing rose to nearly 100% in cells lacking the *NEJ1* (*YLR265C*) gene required for efficient NHEJ, corroborating previous reports². Progressive barcode integration reached near-completion by 11 generations as shown by both PCR amplification (Fig. 2b) and survival on 5-

fluorocytosine, indicating removal of the *FCY1* (*YPR062W*) counter-selectable marker at the barcode locus (Fig. 2c). We observed similar kinetics of barcode integration and guide-donor plasmid self-destruction with a complex pool of guide-donors designed to introduce natural variants (Fig. 2b). Collectively, these results show that precision editing, genomic barcode integration, and guide-donor plasmid self-destruction all reach near-completion by 9 to 11 generations.

Active donor recruitment to breaks improves HR efficiency

Efficient HR is particularly important for multiplexed editing as typical array-synthesis error rates of 1 in 200 mean that 10% of guides should harbor at least one error, impairing target cleavage. Furthermore, guides exhibit variable cleavage efficiencies dependent on intrinsic features of the guide sequence and the target DNA locus^{7,8}. Because cells with functional guide RNAs will undergo cell-cycle arrest during repair, or will not survive editing and undergo cell death^{5,9}, cells containing mutated or low efficacy guides will dominate the population.

We hypothesized that HR efficiency might be limiting for cell survival and sought to enhance HR by active recruitment of the donor to double-strand DNA (dsDNA) breaks. We adapted an endogenous mechanism required for yeast mating type switching from *MATa* to *MATα*¹⁰, where a sequence element called the recombination enhancer (RE) near the *HMLa* donor mediates enhanced HR at the *MAT* locus. *HMLa* donor recruitment requires two interactions: the binding of Fkh1p to the RE, and the recruitment of Fkh1p to the *MAT* locus dsDNA break via binding of the forkhead-associated (FHA) domain of Fkh1p to phosphothreonines on multiple proteins¹¹, including the Mph1p helicase, Fdo1p, and likely additional unidentified proteins¹². Fusing Fkh1p to the LexA DNA binding domain (LexA-Fkh1p) and replacing the RE with LexA sites partially rescues a deletion of the RE¹³. To adapt this mechanism for MAGESTIC, we introduced a tandem array of four LexA sites on the *ADE2* guide-donor plasmid and introduced LexA-Fkh1p on a plasmid harboring constitutive Cas9 (Fig. 3a). We spiked in a plasmid with a nonfunctional guide at 15% to simulate error rates typically observed in oligo libraries. Cells containing a functional *ADE2* guide-donor plasmid, but lacking either LexA-Fkh1p, the LexA sites, or both, were poorly represented, comprising only 8–12% of the surviving colonies (Fig. 3b, Supplementary Table 1). In contrast, the presence of both LexA-Fkh1p and LexA sites led to a >5-fold increase in the percentage of edited colonies in both WT and *nej1* backgrounds (Fig. 3b), with the fraction of red colonies more closely resembling the plasmid input ratios. NGS of the edited locus from the population confirmed that the increase in red colony fraction occurred through an enhancement of HR and not NHEJ (Fig. 3c), demonstrating that active donor recruitment with the LexA-Fkh1p system specifically increases HR efficiency.

Saturation editing to dissect a protein-drug interaction

To validate the high-throughput editing capacity of MAGESTIC, we designed a guide-donor library to saturate a region of the essential eukaryotic gene *SEC14* (*YMR079W*) with amino acid mutations. The Sec14p phosphatidylinositol transfer protein is an attractive drug target in pathogenic fungi¹⁴, and represents the sole essential target of small molecule inhibitors

termed NPPMs (nitrophenyl(4-(2-methoxyphenyl) piperazin-1-yl)methanone)¹⁵. Several mutations that ablate NPPM binding without strongly compromising Sec14p function have been previously identified¹⁵. As this study employed random mutagenesis to select for NPPM-resistant clones, it likely did not test all possible amino acid mutations and also was not capable of identifying mutations resulting in increased NPPM-sensitivity. We reasoned that saturation mutagenesis could provide a complete map of residues important for Sec14p drug interactions.

High-throughput CRISPR editing requires strategies that prevent donor cleavage by the paired guide, while retaining incorporation of the desired variant. Previous approaches have engineered a synonymous mutation in the protospacer adjacent motif (PAM) in addition to the desired variant¹. This strategy has limitations because not all PAMs can accommodate synonymous changes, and because the efficiency of incorporating the desired variant is impaired by greater distance from the PAM¹. Many ORFs also contain regions devoid of NGG SpCas9 PAMs. To circumvent these limitations, we devised a synonymous codon spreading strategy that is robust with respect to such “PAM-deserts” (Supplementary Fig. 3). Our strategy involves spreading synonymous mutations from the target codon towards the Cas9 cut site to prevent donor cleavage and ensure incorporation of the entire edit by limiting the length of microhomology between the Cas9 cleavage site and the target codon. To account for potential effects of synonymous changes we included a synonymous-only donor that left the target codon unchanged. In addition, we introduced each target codon twice using upstream and downstream synonymous changes paired with different guides to control for potential off-target effects (Fig. 4a).

The mutation of essential genes presents a unique challenge: each designed mutation that is not detected in the edited pool could either be an unsuccessful edit or a successful, but functionally detrimental, edit. To resolve these possibilities, we took advantage of previous findings that the lethality of *SEC14* deletion is suppressed by loss-of-function mutations in *KES1* (*YPL145C*) or *CKII* (*YLR133W*) that oppose Sec14p-mediated signaling^{16–18}. Transforming the guide-donor library into a *KES1*-deficient strain harboring WT *SEC14* enabled recovery of detrimental variants and otherwise lethal premature termination codons (PTCs). Mating the edited *SEC14* library to a complementary suppressor strain lacking *CKII* and *SEC14* but containing *KES1* led to (1) the cellular requirement for *SEC14* function and (2) the sole copy of *SEC14* being the edited variant (Fig. 4b). We sequenced the edited window of *SEC14* to assess the counts for each variant, successfully detecting 1361/1382 (98.5%) designed variants at the haploid stage. We found <0.5% of NHEJ-indel events in the sequenced window, consistent with our results on *ADE2* with cells pre-expressing Cas9 from the constitutive *TEF1* promoter (Fig. 3c) as well as previous results⁵. We observed an expected depletion of PTC and proline variants (Supplementary Fig. 4), and generated a profile of functionally important residues, including residues highly intolerant to a large number of non-synonymous changes (Fig. 4c, top panel).

To identify mutations that rendered Sec14p resistant to NPPM inhibition, we grew the diploid *SEC14* library in the presence or absence of sub-lethal doses of the NPPM 4130-1276¹⁵. This approach revealed a rich profile of mutations conferring both resistance and sensitivity to NPPM (Fig. 4c, bottom panel). Replicates of the drug screen revealed high

concordance (Supplementary Fig. 5a), and results with the upstream and downstream synonymous versions of each edit were similar, indicating that the observed phenotypes were due to coding changes and not the synonymous DNA changes that accompanied them (Supplementary Fig. 5b). Our results were also consistent with two previously characterized positions, Y111 and Y122, which frame the Sec14p phosphatidylcholine (PtdCho) head group-coordinating substructure¹⁵. We detected substantial resistance for Y111A and most other Y111 substitutions, with notable exceptions of the Y111Y synonymous control, Y111F, Y111L, Y111I, and Y111M, suggesting that bulky hydrophobic residues at this position stabilize NPPM binding. Previous studies have found that Y122A does not affect sensitivity to NPPMs, while Y122F confers slight resistance¹⁵. Notably, Y122F and Y122W were the only amino acid changes at position 122 which conferred resistance in our assay. To confirm the accuracy of our high-throughput approach, we chose several specific mutations identified by our screen that, relative to the synonymous controls, increased NPPM resistance (A104D, E124R, L126E/C), decreased resistance (A104C, E124M/F) or showed minimal change (E124G, A104V/Y). Recreating these mutants without accompanying synonymous changes and phenotyping them individually revealed excellent concordance with the change in drug resistance indicated by our screen (Fig. 4e, Supplementary Table 2).

Complementing the previously characterized mutations, our approach generated a complete functional map of the Sec14p 102–137 region and its interaction with NPPMs (Fig. 4c). Functional and mutational hotspots fell under four main categories: 1) Positions where most mutations were tolerated for Sec14p function and conferred resistance to NPPM (A104, P108, Y111, H112, D117 and G127); 2) Positions where only a few specific amino acid changes conferred resistance, despite most substitutions having no impact on function (e.g. I103D/F/L/W/Y, Q109C/D/G/W, V121F, E125I/V, L126A/C/D/E/V); 3) Positions intolerant of most changes, while still permissive for NPPM-resistance mutations (P120F/H/M/S/W, E124R); and 4) Positions important for function but harboring no NPPM-resistance alleles (D115, V129). These distinctions are likely the result of a trade-off between preserving the function of the essential Sec14p and interfering with binding of an inhibitor, with some residues having a greater impact on one process than the other. NPPM 4130-1276 docking experiments predicted that, while Y111 and P120 face the Sec14p lipid-binding pocket and could directly interfere with NPPM binding, many of the resistance hotspots (e.g. A104, P108, G127) are far removed from the presumptive NPPM binding site^{14,18}. These substitutions likely impair Sec14p::NPPM interactions in the trajectory by which the NPPM enters the Sec14p lipid-binding pocket, by modulating conformational changes that accompany NPPM entry into the Sec14p lipid-binding pocket, or by changing the conformation of the binding pocket itself.

High-throughput construction of natural variants

A major challenge in quantitative genetics is identifying how individual genetic variants impact phenotype at genome-scale. Engineering variants pertinent to natural populations with MAGESTIC could be used to address this challenge. As proof of principle, we introduced a subset of variants from the well-studied vineyard isolate RM11 into the common laboratory strain S288c. We designed guide-donor pairs to target 30,410 out of

44,020 SNPs, 1,629 out of 3,548 indels, and 3,566 out of 4,754 linked variants (combinations of variants within 5 bp of each other) without the use of accompanying PAM mutations or synonymous changes. These 35,605 variants were selected on the basis of whether they disrupted an NGG PAM or were located anywhere in the 20 bp guide recognition sequence (Fig. 5a). To analyze the dynamics of individual barcoded strains during editing, we compared pre- and post-editing barcode abundances. This comparison can reveal factors affecting guide efficacy, as cells undergoing Cas9-induced double-strand breaks are at a competitive disadvantage. As a control, we first examined barcodes tagging dead guides (guides containing oligo synthesis errors predicted to abrogate target recognition and cleavage) and observed a median enrichment of ~3-fold (Fig. 5b). Enrichment decreased for mutated and near-perfect guides, defined as having mutations with progressively smaller impacts on cleavage efficiency (Fig. 5b). Perfect guides showed median negative fold-changes, consistent with the negative growth effects associated with Cas9 cleavage and subsequent DNA repair at that target site. In addition, we reasoned that some guides would be capable of cleaving both their target locus and the donor, leading to multiple cycles of cleavage and repair at the target locus. To this end, we examined how the median enrichment of sequence-perfect guide-donors behaved as a function of variant distance from the PAM.

As expected, dead guides were enriched 3-fold regardless of variant location. In contrast, cells harboring sequence-perfect guide donors were markedly depleted (Fig 5c). While variants 1 to 10 bp from the PAM showed only mild depletion, variants 11 to 19 bp away exhibited a gradual drop in abundance, with variants at 20 bp from the PAM showing a substantial drop relative to 19 bp. This was unexpected as previous work has suggested that mismatches at the 19th and 20th positions are equally tolerated^{19,20}. Overall, our dataset analyzing the mismatch tolerance of 23,866 distinct guide-donor pairs across the genome suggests that a substantial fraction of donors with SNPs throughout the guide target region may be competent for editing and subsequent resistance to Cas9-guide cleavage.

An improved guide RNA efficacy scoring system for yeast

Even among sequence-perfect guide-donors, we observed a wide range of log fold changes (logFCs) in abundance during editing, suggesting that a subset of sequence-perfect guides are ineffective at target cleavage. To determine whether guides with positive logFCs corresponded to ineffective guides we analyzed the correlation between barcode logFCs and Azimuth efficacy scores, which are widely-adopted machine learning-based scores derived largely from the nucleotide content of the target site and thermodynamics of the guide-target interaction⁸. As expected, we noticed an overall decrease in the distribution of logFC with increasing Azimuth score (Supplementary Fig. 6a). We next tested the effect of PAM sequence on efficacy, and noticed a subtle decrease in effectiveness with guides targeting TGG PAMs, consistent with previous results⁷ (one-sided Wilcoxon test, $p = 6.48E-05$; Supplementary Fig. 6b).

Many of the highest logFC sequence-perfect guide-donors contained poly-T stretches, which we reasoned could promote premature termination of Pol III transcription^{21,22}. We examined each homopolymer by length, observing that T-homopolymers of lengths 3, 4 and 5 were

disfavored more than their A, C, and G-counterparts (Supplementary Fig. 7a). Furthermore, we noticed that T₃ and T₄ resulted in lower efficacy when located at the 3'-end of the guide (Supplementary Fig. 7b). This is likely a consequence of the GTTT sequence in the structural guide component immediately downstream, thus resulting in an extended, imperfect T-stretch. To test whether these imperfect T-stretches can be used to predict guide efficacy, we assigned each guide a score based on the length of the longest imperfect T-stretch, with penalties for interruptions known to reduce Pol III termination²³ (high T-score defined as ≥ 5). Notably, the T-score alone predicted guide efficacy to a similar extent as Azimuth (Spearman $\rho = -0.18$, Pearson $R = -0.19$, both $p < 2.2E-16$ for Azimuth; $\rho = 0.2$, $R = 0.22$, both $p < 2.2E-16$ for T-score). The T-score remained a significant predictor even after accounting for the guide efficacy variance explained by Azimuth score (ANOVA test on Azimuth and T score term, both $p < 2E-16$, Supplementary Table 3). The additional variance explained by the T-score most likely concerns very inefficient guides (T-score ≥ 5 , Fig 5d, Supplementary Fig. 8a), some of which were predicted to be relatively efficient by Azimuth but showed a $\log_{2}FC > 0$ in our dataset. This discrepancy is likely due to Azimuth being trained only on single-, di-, and position-independent nucleotide content⁸, none of which would capture imperfect T stretches.

To confirm that T-scores ≥ 5 are indicative of reduced guide efficacy because of premature Pol III termination, we analyzed RNA levels globally through reverse transcription and targeted sequencing of the HDV-guide-structural RNA transcripts. We normalized guide RNA counts to guide DNA counts and binned by T-score. These results revealed decreasing median guide abundance with increasing T-score, with a significant drop from T-score 4.5 to 5 – the threshold we had defined for high T-scores (Fig. 5e). These results were independent of synthesis-derived errors in the guide, indicating that low RNA levels of high-T score guides are not simply artifacts due to low guide activity (Fig. 5e). As we omitted uninterrupted stretches of six or more T's from our guide designs, all T-scores greater than 5 represent imperfect T-stretches. This suggests that T₅ stretches are more potent terminators than imperfect stretches with T-scores of 5.5 or 6.

Relative to yeast Pol III, mammalian Pol III terminates with shorter T-stretches, including T₄ as well as imperfect stretches such as T₂VT₃^{21,23}. We observed that very few guides in the training set used for the Azimuth algorithm had T-scores ≥ 5 (Supplementary Fig. 8b), which could explain why imperfect T-stretches were not factored in as a predictor. We conclude that incorporation of imperfect T-stretches into machine learning-based models will lead to improved efficacy predictions and superior guide design algorithms for Pol III driven-guides in yeast and likely in higher eukaryotes as well.

Barcodes serve as accurate proxies for edits

To test how well our barcodes reflect their encoded variants, we sequenced the barcodes of 36 clones isolated after editing. We found that 21 contained guide mutations, consistent with the global enrichment of non-functional guides (Fig. 5b), and as expected yielded no edits at the target locus. For the remaining 15, we sequenced the target locus and found 9 WT and 6 donor edits (Supplementary Table 4). Of these 15 clones, 5 exhibited high T-scores ≥ 5 , all of which were WT at the target locus. We therefore estimate an editing efficiency of 6/10

after excluding 5 high T-score guides and 21 mutated guides. We note that due to the enrichment for non-functional guides, the culture size and sequencing depth needed to assay the edited population effectively increases ~5-fold. It is therefore important that the post-editing yeast libraries are not subjected to passage bottlenecks that would result in the loss of low abundance variants. Overall, this work highlights the power of MAGESTIC to rapidly construct thousands of individual genetic variants, constituting a powerful system for rapidly dissecting quantitative traits down to the nucleotide-level by short-barcode sequencing-based counts.

Discussion

Dissecting complex genotype-phenotype relationships has remained a central obstacle in quantitative genetics despite major technological advances in sequencing and genome editing. Assessing the functional impact of genetic variants will be greatly accelerated by robust technologies that can precisely engineer and quantitatively phenotype variants at large scale. In this study, we develop the MAGESTIC platform to engineer single nucleotide and amino acid variants genome-wide and quantify fitness by short barcode sequencing.

MAGESTIC surpasses several limitations of currently available methods, namely the instability of plasmid barcodes and the inability to distinguish between oligo synthesis errors and PCR/sequencing errors in the guide and donor during phenotyping¹⁻³. First, MAGESTIC separates the steps of guide-donor sequence validation from variant quantification by tagging each guide-donor with a unique short (31-mer) barcode during cloning. A single high-throughput sequencing run with 150-bp paired-end reads can associate each unique barcode with a specific guide-donor sequence at the plasmid library stage. Economical, high-throughput phenotyping can then be achieved with 31-bp reads to count each variant without having to sequence the entire guide-donor for each count. In addition, these barcodes can be used to distinguish cells carrying the same guide-donor pair but deriving from independent editing events, providing internal replicates and serving as single-cell tracers. Second, MAGESTIC efficiently integrates the plasmid barcode into the genome and removes residual guide-donor plasmid via plasmid self-destruction. Integration of the barcode offers several advantages: (1) phenotyping is not confined to environments requiring marker selection, (2) each cell harbors only a single barcode rather than a variable copy number plasmid, and (3) thousands of individual strains can be readily isolated and identified *en masse* from a mutant pool using recombinase-directed indexing²⁴. This allows downstream validation of individual variants as well as spatially-separated phenotyping, such as measuring productive capacity for bioengineering or protein localization in high-throughput microscopy.

While a previously published guide-donor method developed in prokaryotes (CREATE) employed a 1-step cloning procedure by including the guide RNA promoter between the donor and guide sequence¹, this method is not amenable to eukaryotic systems as no eukaryotic promoters are short enough to be included given the current length limitations of array-based oligonucleotide synthesis. A second cloning step is required to either insert the guide RNA promoter, or the structural guide component, with the downside of potentially introducing bias into the library. By maintaining very high coverage at the first step of

cloning (a mean of >20 barcodes per variant), we demonstrate that we can maintain complexity and uniform representation of variants in the library (Supplementary Fig. 1). In addition, we and others have found that selectable markers in the inserts for the second cloning step remove undesirable background².

One of the central challenges in precision genome engineering is creating desired changes with high fidelity and efficiency while avoiding competing pathways of NHEJ-indels and cell death. To address this challenge, we developed a method to actively recruit the donor DNA to the site of DNA breaks using a hybrid LexA-Fkh1p fusion system, and demonstrated >5-fold increase in HR efficiency. Active donor recruitment prevents cells with non-functional guides from overtaking those with functional guides, enabling improved representation of engineered genetic variants in library-scale editing. Although others have shown that tethering donor DNA to Cas9 promotes increased HR^{25–27}, these approaches are not amenable to high-throughput screening as the guide RNA and donor DNA must be expressed separately prior to physically associating with Cas9. Recruitment of donor DNA to Cas9 breaks by the Fkh1p-phosphothreonine-mediated mechanism offers additional advantages over direct tethering to Cas9, as multiple copies of donor can be recruited to the break, and enhanced repair does not depend on persistence of Cas9 association with the break. As FHA-recruitment to dsDNA breaks is conserved from yeast to humans²⁸, it is likely that this mechanism can be adapted to improve editing in NHEJ-prone mammalian systems. A previously published guide-donor method developed in prokaryotes (CREATE) demonstrated significant toxicity due to editing resulting in ~5% survival¹, which is on the order of the ~10% survival we show for yeast in the absence of the Fkh1p-LexA fusion system (Fig. 3b). Active donor recruitment should therefore improve library-scale editing approaches in bacterial systems as well.

A major challenge for engineering SNPs is the high-degree of sequence similarity between the guide and the donor, as recognition and cleavage of the donor DNA will result in loss of the variant through cell death or mutation by NHEJ. A second challenge is the availability of PAMs near the SNP, as successful incorporation of the SNP by HR decreases with increasing distance from the cut site. In this study, we use WT Cas9 and thousands of guide RNAs across the genome, and find that SNPs can be tolerated to differing extents along the guide region, with a significant drop from the 19th to 20th bp positions from the PAM. Ultimately, engineered variants of Cas9 exhibiting reduced mismatch tolerance but maintaining high on-target activity will aid in successful engineering of SNPs throughout the guide region^{29–32}. Lastly, we demonstrate that Pol III-terminating T-stretches play a substantial role in dictating guide efficacy in yeast. The use of different promoters, such as Pol II promoters with ribozymes to release the guide from the 5′-cap and poly(A) elements, may address this inherent limitation of delivering guides from the Pol III promoter to T-rich genomic targets. Furthermore, accommodating RNA-guided nucleases with different PAM preferences will broaden the target space of the MAGESTIC system, while specific targeting of highly repetitive regions will remain a challenge with all RNA-guided nuclease approaches. Overall, MAGESTIC enables tens of thousands of specific genetic variants across the genome to be created in a manner that is compatible with robust phenotyping across hundreds of conditions, and will significantly advance our understanding of the genotype-environment-phenotype relationship.

Methods

Yeast strains and media

The yeast strain background used in all experiments is a derivative of BY (S288c) named DHY214 (*MAT α his3 1 leu2 0 ura3 0 lys2 0*) in which genetic defects have been repaired to improve sporulation [*MKT1(30G) RME1(INS-308A) TAO3(1493Q)*] and mitochondrial genome stability [*CAT5(91M) MIP1(661T) SALI⁺ HAPI⁺*]. To generate the landing pad at the chromosomal barcode locus, this strain was first transformed with pKR76 (P_{TEF1}-Cas9 with *URA3* and *hphMX* markers; <https://benchling.com/s/pregddyA>) to yield yKR15. yKR15 was then transformed with V79 (*FCY1* guide driven by the tRNA(Tyr)-HDV ribozyme promoter³³; <https://benchling.com/s/1M0BfuaJ>) and a linear donor constructed by annealing and extended overlapping oligonucleotides oKR86-oKR87, which introduced a precise deletion in the *FCY1* open reading frame. A control experiment with an irrelevant donor targeting *CANI* yielded no surviving colonies, confirming dependence of cell survival on HR via donor DNA. All 8 clones examined exhibited the correct deletion of *FCY1* as confirmed by PCR of the locus and by growth on 5-fluoro-cytosine (5-FC). One clone was selected and named yKR26. To generate the chromosomal barcode locus, the SCEI-*FCY1*-SCEI landing pad was amplified from yACJ2, along with primers with 50 bp of homology upstream and downstream of the SCEI sites to enable integration of the guide-donor cassettes. The homology sequences were randomly generated using the python 2.7 module *random*, and checked for lack of homology to the yeast genome by BLAST³⁴. The downstream integration sequence was followed by the *URA3* promoter and the first half of the *URA3* gene, followed by half of an artificial intron and the lox71 site, yielding yJS4 (<https://benchling.com/s/seq-8KWFCuPiwZUbhrPRqxFe>). The latter construct was included to render these strains compatible with recombinase-directed indexing (REDI)²⁴. Transformation of plasmid libraries was performed with standard lithium-acetate/PEG/ssDNA procedure³⁵.

Plasmid design

We designed guide-donor plasmids with the precision editing guide under control of a tRNA(Tyr)-HDV promoter. For the *SECI4* editing we used pKR216 (<https://benchling.com/s/seq-eyZHbUi3B7xm2BGy0Lbm>), which is a 2 μ -plasmid containing counter-selectable *FCY1*, a site for guide X cleavage (TAGGGATAACAGGGTAATGGtgg, PAM in lowercase), and a tandem array of four LexA-sites as well as upstream and downstream homologies for barcode integration. For the natural variant experiments, we created pKR348 (<https://benchling.com/s/seq-jGc3L4hiMsI7PFs3wtcg>), a 2 μ -plasmid which contains extended overlaps to the barcoding locus, the LexA-Fkh1p fusion under control of the *ADHI* promoter, a tandem array of four LexA sites, a guide X cleavage site, and 200 bp upstream and 300 bp downstream homologies to the barcoding locus. For the barcoding guide (guide X) we tested three different promoters, *RPR1(TetO)*²⁴, *SNR52*, and tRNA(Tyr)-HDV. As all three promoters showed similar levels of barcode integration and plasmid destruction (Supp. Fig. 1), *RPR1(TetO)* was chosen to drive guide X on pKR348 to enable the option of TetR-controlled expression. For the natural variant experiments, Cas9 was expressed from pKR291 (<https://benchling.com/s/seq-tA9exl8LT94qdspLOF2b>), under

the control of a galactose-inducible promoter to allow for temporal control of Cas9 expression.

Analysis of editing, barcoding, and plasmid removal kinetics

For experiments described in Figure 2a and 2b, cells were cultured in 48-well plates in an Infinite plate reader (Tecan) at 30°C with orbital shaking. OD₆₀₀ was followed by taking measurements every ~15 minutes. Cultures were maintained in log phase growth by passaging cultures every 2 doublings, when an aliquot of the culture was additionally transferred to a collection plate at 4°C (Torrey Pines) for further processing. The sub-passage and culture sampling steps were triggered by a pre-defined OD (0.6), not by time elapsed. Liquid transfers were performed automatically using a Freedom EVO liquid handling system (Tecan), which was controlled by custom Pegasus software (Tecan). For the colony count analysis for survival on 5-fluoro-cytosine (5-FC) versus YPD, a strain harboring the RM11 natural variants library was grown in quadruplicate in CSM-URA-HIS+galactose from OD 0.05 to OD 1.6 for the initial 5 generation time point and sub-passaged into fresh CSM-URA-HIS+galactose at OD 0.05 for subsequent time points. At the indicated generations, ~1000 cells were plated and the number of colonies on YPD and 5-FC were manually counted. All editing libraries were maintained in CSM-URA-HIS+glucose prior to galactose induction.

Active donor recruitment by LexA-Fkh1p

We cloned LexA-Fkh1p under control of the *ADHI* promoter into pKR76 (<https://benchling.com/s/pregddyA>), a pRS416-based vector also containing Cas9 under the *TEF1* promoter, to give pKR193 (<https://benchling.com/s/WLoXhBjL>). pKR76 and pKR193 plasmids were separately transformed into yJS4 and an *nej1* null version of yJS4 (yKR139). We then made two mixes of plasmids. The first mix contained 85% by mass an *ADE2* guide-donor 2 μ -plasmid without LexA sites (pKR184), and 15% a 2 μ -plasmid without a functional guide (pKR185). The second mix contained 85% by mass an *ADE2* guide-donor 2 μ -plasmid with 4 LexA sites (pKR194; <https://benchling.com/s/ozgmJR2v>), and 15% a 2 μ -plasmid without a functional guide (pKR185; <https://benchling.com/s/MJO8mPTq>). These mixes were transformed using lithium acetate transformation into the four strains expressing Cas9 with or without Fkh1p. The colonies were allowed to grow for a week and then colony counts were generated by counting sectors of the plate to give relative counts for edited colonies, and then plates were washed and gDNA extracted from the population and sequenced at the *ADE2* locus (see Supplementary Table 1).

Analysis of editing outcomes at *ADE2* and *SEC14* loci

The edited regions for *ADE2* and *SEC14* were amplified with Illumina adapters and sequenced with MiSeq v2 2 × 150 bp reads. All reads were processed with the following BBtools commands with default settings (sourceforge.net/projects/bbmap/). Reads were trimmed with *bbduk* (version 37.17), merged with *bbmerge* (version 37.17), and mapped to reference files containing the WT and designed variant sequences using *bbmap* (version 37.17). Reads mapping with an insertion or deletion in the guide target or PAM sequence were designated as NHEJ-indel events, while reads mapping imperfectly to designed variant

sequences in the region harboring the sequence changes were designated as imperfect donor repair events using custom python scripts (see Code Availability).

Guide-donor library design

For *SEC14* saturation mutagenesis, the guide-donor oligonucleotide sequences encoded mutations to convert each amino acid to the other 19 amino acids as well as a stop codon. The highest frequency codon for each amino acid was utilized for each target amino acid change. For each amino acid, the nearest upstream and nearest downstream PAMs were located and their corresponding guides selected. For the donor DNA, synonymous codons (selected on the basis of the largest hamming distance relative to the codon, with the exception of suboptimal codons with usage frequencies less than 10%) were introduced between the target amino acid and the Cas9 cut site (3 bp upstream of the PAM), until a disruption score of 6 was achieved for the synonymous-only donor control. Disruption scores were calculated by aligning the guide to the donor, with disruptions in the GG of the NGG PAM counting as 3 for each disruption, disruptions in the PAM proximal 10bp (i.e. “seed” region) as 2 each, and disruptions in the PAM distal 10bp as 1 each. Disruptions refer to either mismatches or indels in the alignment. The disruption score of 6 was intended to ensure complete lack of guide cleavage activity on the donor DNA. For the natural variant libraries, the guide-donor oligonucleotide sequences were designed by first generating VCF files by comparing bam files from novoalign mapping (version 3.07.00, default settings) of Nextera-prepped whole-genome sequencing samples (75 bp paired-end reads) for RM11-1a and SK1 against DHY214 with SICtools³⁶. For each entry in the VCF file, all combinations of variants within 5 bp were included in a “linked” variant category to account for amino acid changes and enable construction of multi-nucleotide variants. Each variant was analyzed on the basis of disrupting either an NGG PAM or the 20 bp upstream of an NGG PAM. Guide RNAs or donor DNAs harboring restriction sites used in the cloning steps (NotI, AscI, or BspQI) were removed from the design. For all libraries, guides were disqualified if they contained the canonical Pol III terminator T₆. The BspQI site with an overhang enabling ligation of the structural guide RNA (GTTTAgagagc, restriction site recognition sequence in lowercase) was inserted in between the guide and the donor; a forward priming site (GGACTTTggcgcgcc) was appended just upstream of the guide sequence; and 15 bp serving as subpool-specific priming sites were appended to the 3′-end (just downstream of each donor) for each oligo sequence.

Barcoded guide-donor library cloning

Array-synthesized guide-donor oligos were obtained from Twist Biosciences (RM11 library) or Agilent Technologies (SK1 library) at the 170-mer scale. We amplified subpools with a forward primer harboring an AscI restriction site at its 3′-end and a reverse primer with a NotI site at its 5′-end followed by a degenerate barcode encoding a pseudo-random sequence (either NNNVHTGNNNVHTGNNNVHTGNNNVHTGNNN or NNNTG VHNNTGVHNNNTGVHNNNTGVHNNN) that excludes illegal restriction sites (NotI, AscI, and BspQI), followed by subpool-specific priming sequence. The guide-donor oligos were amplified using KAPA HiFi polymerase as directed by the manufacturer in 50 uL total reaction volume with an initial denaturation of 98°C for 1 min, and then 15 cycles of 98°C 10s, 60°C 20s, and 72°C 30s. Reactions were column-cleaned with the Qiagen

QIAquick PCR purification kit. NotI and AscI sites enable sticky end cloning into a multi-copy recipient vector, with the AscI site at the 3'-end of the guide RNA promoter. 5 ug of each PCR-cleaned reaction was cut with 2 uL of AscI (NEB) and 2 uL of NotI (NEB), 10 uL of 10× CutSmart buffer (NEB), and incubated at 37°C for 1 hour followed by 20 minutes of heat inactivation at 80°C. Reactions were column-cleaned and 400 ng of each insert was ligated with T4 DNA ligase into 1 ug of recipient vector (>7:1 insert:vector) treated with NotI, AscI, as well as CIP (NEB) – either pKR216 (*SEC14* library) or pKR348 (natural variants library) – in a total volume of 20 uL. Ligation reactions were ethanol precipitated by adding 80 uL 100% EtOH and 2 uL of 5M NaOAc pH 5.2 with 1 uL of glycoblue (Ambion), incubated on ice for 10 minutes and spun at 13.2 krpm for 5 min, washed with 70% ethanol, and then re-suspended in 3 uL of nuclease-free water (IDT). 1 uL of each reaction was then electroporated into 20 uL NEB 10-beta in 0.1 cm-gap electroporation cuvettes (Bio-Rad) with the Bio-Rad GenePulser electroporator using the settings 1.7 kV, 200 Omega, and 25 μF. Typical time constants ranged from 4.5–4.8 milliseconds. Cells were recovered for 1 hour at 37°C in pre-warmed SOC and plated onto pre-warmed LB+Carb plates, with a 1:1000 dilution to get estimated colony counts. Typical colony counts on this dilution plate ranged from 200 to 2000. The following day cells were scraped from the plates, and plasmids were extracted with the Qiagen miniprep kit. The guide and donor sequences are separated by a type IIS restriction site (BspQI) that enables cloning with an arbitrary overhang, in this case the GTTT directly 3' of the guide sequence, to enable cloning in the constant structural component of the guide RNA. 5 ug of the plasmid library was cut with 2 uL of BspQI and 2 uL CIP in a total volume of 100 uL, and column-cleaned. The insert containing the structural guide RNA component with yeast-specific (e.g. *HIS3*) and bacterial-specific (e.g. *kanR*) selection markers was amplified from pKR340 (<https://benchling.com/s/seq-7PTZ8FoBXCNwIuXNHSHL>) with primers harboring BspQI sequences at their 5'-ends. The reverse primer included an additional barcode (bc*; either NNNNNN or NNNNNNHVVNHHBBHBHD) situated 3' of the Illumina read 2 priming sequence, modified to contain a G-to-A SNP at the first position of the BspQI site. These second-step libraries were ligated and electroporated with the same conditions described above (i.e. same as the first step libraries), except that the bacteria were selected with kanamycin to enable enrichment of vectors harboring the insert.

SEC14 mutagenesis and phenotyping

SEC14 is an essential gene. To detect mutations that impair *SEC14* function without causing cell death, we took advantage of two known 'Sec14p bypass' suppressors, *CKI1* and *KES1* (Cleves et al., 1991; PMID: 1997207). We introduced all *SEC14* genetic modifications in a *MATa kes1* haploid strain carrying the plasmid pKR197 (<https://benchling.com/s/s3Xpa5CQ>) expressing Cas9 from the *TEF1* promoter and LexA-Fkh1p from the *ADH1* promoter. We also created a second suppressor strain by deleting the entire *SEC14* open reading frame (ORF) in a *MATa cki1* haploid strain. Following mutagenesis of *SEC14* using our CRISPR/Cas9 editing system, the *MATa sec14* mutant pool was mated *en masse* to the *MATa cki1 sec14* suppressor strain, by mixing equal numbers of *MATa* and *MATa* cells in 3mls YPD, and incubating that mixture for 6 hours at 30°C with moderate shaking. Diploids were selected by plating the mated culture on media lacking methionine and lysine. After 2 days of growth at 30°C, diploid colonies were washed off the plate with

water, and aliquots were archived at -80°C in 25% glycerol. The resulting diploid pools contained strains whose viability were dependent on a single copy of *SEC14* containing a genetic modification introduced by our guide-donor library (i.e. *MAT α* , *sec14* Δ *SEC14*^{*}, *cki1* Δ *CKII*, *KES1*/*kes1*).

To phenotype our library of *SEC14* variants, we used competitive growth followed by Illumina sequencing of the edited locus to quantify individual strain fitness. *SEC14* variant pool cultures were inoculated from frozen aliquots to a final concentration of 0.1 OD/ml in 20ml of YPD medium, and grown for 4 hours at 30°C with moderate shaking. 700 μl aliquots of this culture were then transferred to 48-well plates and grown in the presence of $8\mu\text{M}$ of the NPPM 4130-1276, or DMSO as a control. Each condition was represented by duplicate cultures. These 48-well plates were grown in an Infinite plate reader (Tecan) at 30°C with orbital shaking, which allowed growth of cultures to be continuously monitored by taking OD₆₀₀ measurements every ~ 15 minutes. Cultures were maintained in log phase growth by automated passaging, in which 80 μl of culture was transferred to a new well containing 620 μl of media upon reaching a ‘trigger’ OD of 0.76. Liquid transfers were performed using a Freedom EVO liquid handling system (Tecan), which was controlled by custom Pegasus software (Tecan). After 3 passages (~ 12 generations total growth), 600 μl of culture at OD 0.76 was transferred to a collection plate stored at 4°C (Torrey Pines) for further processing. Genomic DNA was extracted from saved cells using the Yeastar genomic kit, as well as an equivalent number of cells from the edited haploid pool, and the diploid “time zero” pool.

From each of these samples, the edited region of *SEC14* was then amplified by PCR containing adapters for Illumina sequencing (NextSeq). Paired-end reads were quality trimmed by bbdduk and then merged by bbmerge. Merged read counts mapping to each allele were enumerated by searching for perfect matches to the designed donors. A pseudocount of 1 was added to the number of reads assigned to each variant in each sample. Variant read counts observed in the diploid “time zero” pool were used generate the *Relative Variant Abundance* heatmap in Figure 4c. To calculate *Log2 NPPM Resistance* for each variant, read counts for each duplicate sample were first averaged, and then a log₂ ratio of the NPPM-treated and -untreated cultures was calculated [i.e. $\text{Log}_2(\# \text{ reads } + \text{NPPM} / \# \text{ reads } - \text{NPPM})$]. To center the data, we calculate the average log₂ ratio of 44 synonymous *SEC14* control variants, and subtracted that value (-1.848) from all other log₂ ratios. These numbers were used to generate the *Log2 NPPM Resistance* heatmap in Figure 4c. Variants which garnered fewer than 10 reads in each of the samples were excluded from this plot. In cases where the same mutation was represented by multiple variant strains (e.g. upstream and downstream synonymous versions), the average *Log2 NPPM Resistance* was used to color the heatmap.

To validate the NPPM resistance results, we generated 11 *SEC14* variants individually in a WT background to confirm the accuracy of our suppressor strategy and retested their resistance to 4130-1276 in pure cultures. These variants (A104D, A104V, A104Y, A104C, E124R, E124G, E124M, E124F, L126E, L126C, and L126I) were selected because they exhibited a range of NPPM resistance phenotypes (Figure 4). Briefly, the strain DHY214 was transformed with a plasmid expressing both constitutive Cas9 and a guide RNA directed to the *SEC14* locus, in the presence of 11 different double-stranded DNA donors encoding the desired mutation surrounded by 60 bases of homology to *SEC14*. Notably, synonymous

changes were not introduced in these variants. Introduction of the desired mutation was confirmed by Sanger sequencing. Multiple independent clones (2–4) for each variant, plus empty vector (EV) controls were cultured to saturation overnight in YPD liquid media, diluted to OD 0.1 the next day, and grown in 100 μ l cultures in a 96-well plates, either in the absence or presence of 8 μ M 4130-1276. Growth in each well was monitored in a GENios plate reader (Tecan) by taking OD600 measurements every ~15 minutes for the duration of the experiment (~20 hours). Data from representative wells are plotted in Figure 4. All OD measurements are provided in Supplementary Table 2.

Protein Preparation, Homology Modeling and Computational Docking

Protein Preparation—The X-ray crystal structure of Sec14p (PDB ID 1AUA)³⁷ was obtained from PDB repository (www.rcsb.org). The protein models were prepared using the *Protein Preparation Wizard* panel in the Schrödinger suite (2017-4, Schrodinger, LLC, New York, NY, 2017). Complete structure of Sec14p was optimized with the OPLS_2005 forcefield in the Schrödinger suite to relieve all atom and bond strains found after adding all missing side chains and/or atoms. The small molecule model structure for compound 4130-1276 was prepared and energy minimized in MOE (2016.08; Chem. Comp. Group Inc., Montreal, Canada) and the lowest energy conformation was selected for docking.

Homology Modeling—A homology model for the closed conformer of Sec14p was generated using the MOE suite (2016.08; Chem. Comp. Group Inc., Montreal, Canada) based on the templates of the open conformer of Sec14p (PDB ID 1AUA)³⁷ and the closed conformer of Sfh1p bound to PtdIns (PDB ID 3B7N)³⁸. Gate residues in the Sec14p open conformation (I215 – Y247) were removed from that template structure prior to modeling whereas the corresponding gate residues in the closed conformation in Sfh1p/PtdIns were retained. In addition, residues Ala 84 – Gln 111 on the far side of the binding pocket from the gate were removed from the Sfh1p template prior to modeling since they were structurally divergent from the corresponding Sec14p residues. By default, ten independent intermediate models were generated. These different intermediate homology models were generated as a result of permutation selection of different loop candidates and side chain rotamers. The intermediate model, which scored best according to the Amber99 forcefield, was chosen as the final model and was then subjected to further optimization.

Computational Docking—Computational docking was carried out using the genetic algorithm-based ligand docking program GOLD 5.2.1³⁶. GOLD explores ligand conformations fairly exhaustively and also provides limited flexibility to protein side chains. For computational docking crystal structure of Sec14p in an open conformation (PDB ID 1AUA) and homology model in closed conformation was used. The active site was defined by taking residue Ser173 in the crystal structure as a reference center to define protein binding site of radius 6 Å around it, with the GOLD cavity detection algorithm. GOLD docking was carried out without using any constraints or biases to explore all possible diverse solutions. In order to explore all the possible binding modes, docking was carried out to generate diverse solutions with early termination turned off. All other parameters were as the defaults. Compound 4130-1276 was then docked and scored using CHEMPLP scoring function within GOLD as it has been found to give the highest success rates for both pose

prediction and virtual screening experiments against diverse validation test sets. Ligand was docked in independent runs with early termination of ligand docking was switched off, and top 3 best solutions were retained for each run.

Evaluating library representation

In order to assess changes in library representation from the initial oligo library through the transformation into yeast, coverage of barcodes and designed guide-donor variants (features) was compared across the different stages of library construction (Supplementary Fig. 1). Guide-donor cassettes were amplified with custom-designed indexed primers containing Illumina read primer sequences and sequenced on Nextseq 550 v2 150×150 bp paired-end format (step 1 plasmids), 31×45 bp paired-end format (step 2 plasmids) or 31×120 bp paired-end format (yeast libraries). For determining variant representation in the initial oligo library, 134 bps of the guide-donor sequence were extracted from the forward reads and mapped to the designed variant library using BLASTn alignment. Alignments with greater than 98% identity for length \geq 133 bps were used to determine the number of guide-donor variants represented. To examine step 1 coverage from paired-end reads, barcodes were extracted from the first 31 bp of the forward read; all reads for a given barcode were then collapsed to generate a guide-donor consensus sequence for mapping to the library reference using BLASTn. Reads in subsequent steps were mapped directly to the library reference using step 1 annotations.

Pre- and post-editing dynamics

For the analysis in Fig. 5 and Supplementary Fig. 5, 6, 7 and 8, the RM11 natural variants yeast library was recovered from a glycerol stock in SC-URA-HIS Glucose medium (6.9g/l yeast nitrogen base (Formedium), 2% D-glucose (Sigma), 1.92g/l -URA-HIS dropout mixture (Formedium)) for 2h, then washed and transferred to SC-URA-HIS Galactose medium (2% galactose instead of glucose) for 7 generations of editing. For editing, the recovered stock was split into 5 replicates, each inoculated at $OD_{600} = 0.00327$, corresponding to an $\sim 1000\times$ coverage of the library. Generations were counted based on OD_{600} at the time of sampling. Genomic DNA was extracted using the MasterPure kit (Epicentre) and custom-designed indexed primers containing Illumina read primer sequences were used to amplify the barcodes. Samples were sequenced with Nextseq 550 v2 31×120 bp paired-end format and barcode counts derived by mapping the 31-mer barcode read to the step 1 reference table. For analysis of barcode dynamics during editing, barcode counts were filtered to remove barcodes not present in the pre-editing sample and barcodes missing in more than one of the five post-editing samples. We further required a minimum count of 20 for barcodes pre-editing. We used edgeR to obtain normalized counts and determine fold-change for each barcode during editing. For the analysis in Figure 5c, 5d, Supplementary Fig. 6, 7, 8 we only included barcodes tagging a dead guide or sequence perfect guide-donors. We further excluded all barcodes tagging sequence perfect guide-donors aligning to other parts of the genome with less than two mismatches. T-score is defined by length of longest T stretch in the guide (with the downstream sequence GTTT) with up to two non-T residues, with penalties of 0.5 for one non-T residue and 2.5 for two non-T residues. A high T-score is defined as ≥ 5 , based on the median log fold change of these score bins being > 0 . For Fig. 5d, original T-scores were re-binned into five bins to

group original score bins with similar log fold-change distributions and thereby remove redundancy (Bin 0: 0 – 1.5, Bin 1: 2 – 3.5, Bin 2: 4 – 4.5, Bin 3: 5 – 6.5, Bin 4: 7.5 – 9.5). To allow visual comparison between the discrete T-score and the continuous Azimuth score in Fig. 5d, we also binned the Azimuth score, such that the same number of barcodes was included in each bin as for the T-score. We ordered barcodes according to decreasing Azimuth score and assigned the first n barcodes (where n = number of barcodes in T-score bin 0) to Azimuth bin 0, and so on for the rest of the barcodes. For Fig. 5d, Supplementary Figures 6b, 7b we used a one sided Wilcoxon rank-sum test to determine if the difference in location between groups was greater than 0. ANOVA terms and parameters are given in Supplementary Table 3.

Barcode to edit correspondence

Genomic DNA of 36 individual colonies was extracted using the MasterPure kit (Epicentre), and the barcode locus was amplified and sent for Sanger sequencing. We matched the 31-mer barcodes to our step1 reference tables and used the CIGAR strings in the reference tables to mark guides containing a mutation relative to the design (number of perfect matches at beginning of CIGAR < 20). To confirm these guides were indeed mutated we amplified the guide sequences from the barcode locus separately and used BLAST+ (version 2.4.0) to align these traces to the oligo library designs. We also extracted the donor sequences from the Sanger traces and aligned them to the oligo library designs to ensure that only the designed mutations were encoded in the donor. We designed primers for amplifying the target sites using primer3 (release 2.3.7), such that the forward and reverse primers were located symmetrically around the expected edit and the final product size would be 550 – 600bp. We further specified a maximal GC content of 60%, length between 18 and 25nt, and the presence of at least one G or C at the 5' and 3' ends of each primer. The target sites were amplified using the previously extracted genomic DNA and sent for Sanger sequencing. To determine editing outcome we aligned the Sanger traces to the yeast genome (R64-2-1) using BLAST+.

Statistical analysis, reproducibility, and software

Detailed information summarizing the experimental design, statistical parameters, and materials and reagents can be found in the accompanying Life Sciences Reporting Summary.

All statistical analyses were performed in R³⁹ using the stats package (version 3.3.3 and 3.5.0), with the numbers tested indicated in the main or supplementary figures. Changes in barcode dynamics were analyzed using the *edgeR* package^{40,41} (version 3.16.5). One-sided Wilcoxon rank sum tests for group comparisons were performed using the *wilcox.test()* function, correlation was estimated using the *cor.test()* function and the ANOVA analysis used the *lm()* and *anova()* functions. Box plot elements show the median (black line) and quantile values (box denotes 25th and 75th quantile), with outliers shown as black dots outside of the box whiskers. Violin plots show median (black dot), 25th and 75th quantile (black line) and distribution of the groups. For Figure 5c, the number of barcodes per group is given below. For each PAM distance, the first value corresponds to the number of barcodes tagging dead guides and the second value to number of barcodes tagging perfect-guide-donors. $N_{20} = 101/487$, $N_{19} = 74/562$, $N_{18} = 89/743$, $N_{17} = 101/789$, $N_{16} = 88/747$,

$N_{15} = 117/741$, $N_{14} = 125/951$, $N_{13} = 94/823$, $N_{12} = 118/915$, $N_{11} = 119/973$, $N_{10} = 123/1158$, $N_{09} = 118/1274$, $N_{08} = 137/1498$, $N_{07} = 112/1373$, $N_{06} = 125/1855$, $N_{05} = 126/1799$, $N_{04} = 114/1544$, $N_{03} = 116/1572$, $N_{02} = 152/1832$, $N_{01} = 121/1677$, $N_{00} = 4/61$, $N_{-01} = 29/297$, $N_{-02} = 25/330$.

All figures were prepared using Adobe Illustrator CS6. Plots were generated in R using package ggplot2⁴² (version 2.2.1) or Python 2.7 or 3.6.3 using matplotlib⁴³ and seaborn⁴⁴ plotting libraries. The heatmaps in Fig. 4 were generated with Spotfire (version 7.6.1). All analyses in Fig. 5 and Supplementary Fig. 5, 6, 7 and 8 were performed in R³⁹ (version 3.3.3) and plots were generated using ggplot2⁴².

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants from the US National Institutes of Health (P01HG000205 to L.M.S. and R.W.D., R01GM121932-01A1 to R.P.S., U01GM110706-02 to R.W.D., RO1GM61766 to J.E.H., and RO1GM44530 to V.A.B.), the National Institute of Standards and Technology (70NANB15H268 to M.L.S), and the European Research Council Advanced Investigator Grant (AdG-294542 to L.M.S). K.R.R. was supported by a National Research Council postdoctoral fellowship. A.T. and V.A.B. were supported by the Robert A. Welch Foundation (award BE-0017). S.C.V. was supported by a Swiss National Science Foundation postdoctoral fellowship (P2EZP3_165220). Certain commercial equipment, instruments or materials are identified in this document. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose. We thank the EMBL Genomics Core Facility for support and optimization of barcode sequencing protocols. This work is dedicated to the memory of Joe Horecka (12/1/1963-10/20/2017).

References

1. Garst AD, et al. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat Biotechnol.* 2017; 35:48–55. [PubMed: 27941803]
2. Sadhu MJ, et al. Highly parallel genome variant engineering with CRISPR/Cas9 in eukaryotic cells. *bioRxiv.* 2017; 147637doi: 10.1101/147637
3. Guo X, et al. High-throughput creation and functional profiling of eukaryotic DNA sequence variant libraries using CRISPR/Cas9. *bioRxiv.* 2017; 195776doi: 10.1101/195776
4. Michlits G, et al. CRISPR-UMI: single-cell lineage tracing of pooled CRISPR-Cas9 screens. *Nat Methods.* 2017 advance online publication.
5. DiCarlo JE, et al. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Res.* 2013; 41:4336–4343. [PubMed: 23460208]
6. Ugolini S, Bruschi CV. The red/white colony color assay in the yeast *Saccharomyces cerevisiae*: epistatic growth advantage of white *ade8–18*, *ade2* cells over red *ade2* cells. *Curr Genet.* 1996; 30:485–492. [PubMed: 8939809]
7. Doench JG, et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol.* 2014; 32:1262–1267. [PubMed: 25184501]
8. Doench JG, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol.* 2016; 34:184–191. [PubMed: 26780180]
9. Clikeman JA, Khalsa GJ, Barton SL, Nickoloff JA. Homologous recombinational repair of double-strand breaks in yeast is enhanced by MAT heterozygosity through yKU-dependent and -independent mechanisms. *Genetics.* 2001; 157:579–589. [PubMed: 11156980]
10. Wu X, Haber JE. A 700 bp cis-acting region controls mating-type dependent recombination along the entire left arm of yeast chromosome III. *Cell.* 1996; 87:277–285. [PubMed: 8861911]

11. Sun K, Coïc E, Zhou Z, Durrens P, Haber JE. Saccharomyces forkhead protein Fkh1 regulates donor preference during mating-type switching through the recombination enhancer. *Genes Dev.* 2002; 16:2085–2096. [PubMed: 12183363]
12. [Accessed: 2nd March 2018] Binding of the Fkh1 Forkhead Associated Domain to a Phosphopeptide within the Mph1 DNA Helicase Regulates Mating-Type Switching in Budding Yeast. Available at: <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1006094>
13. Li J, et al. Regulation of Budding Yeast Mating-Type Switching Donor Preference by the FHA Domain of Fkh1. *PLOS Genet.* 2012; 8:e1002630. [PubMed: 22496671]
14. Chayakulkeeree M, et al. SEC14 is a specific requirement for secretion of phospholipase B1 and pathogenicity of *Cryptococcus neoformans*. *Mol Microbiol.* 2011; 80:1088–1101. [PubMed: 21453402]
15. Nile AH, et al. PITPs as targets for selectively interfering with phosphoinositide signaling in cells. *Nat Chem Biol.* 2014; 10:76–84. [PubMed: 24292071]
16. Fang M, et al. Kes1p shares homology with human oxysterol binding protein and participates in a novel regulatory pathway for yeast Golgi-derived transport vesicle biogenesis. *EMBO J.* 1996; 15:6447–6459. [PubMed: 8978672]
17. Li X, et al. Analysis of oxysterol binding protein homologue Kes1p function in regulation of Sec14p-dependent protein transport from the yeast Golgi complex. *J Cell Biol.* 2002; 157:63–77. [PubMed: 11916983]
18. Cleves AE, et al. Mutations in the CDP-choline pathway for phospholipid biosynthesis bypass the requirement for an essential phospholipid transfer protein. *Cell.* 1991; 64:789–800. [PubMed: 1997207]
19. Fu BXH, St Onge RP, Fire AZ, Smith JD. Distinct patterns of Cas9 mismatch tolerance in vitro and in vivo. *Nucleic Acids Res.* 2016; 44:5365–5377. [PubMed: 27198218]
20. Hsu PD, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol.* 2013; 31:nbt.2647.
21. Arimbasseri AG, Rijal K, Maraia RJ. Transcription termination by the eukaryotic RNA polymerase III. *Biochim Biophys Acta.* 2013; 1829:318–330. [PubMed: 23099421]
22. Braglia P, Percudani R, Dieci G. Sequence Context Effects on Oligo(dT) Termination Signal Recognition by *Saccharomyces cerevisiae* RNA Polymerase III. *J Biol Chem.* 2005; 280:19551–19562. [PubMed: 15788403]
23. Orioli A, et al. Widespread occurrence of non-canonical transcription termination by human RNA polymerase III. *Nucleic Acids Res.* 2011; 39:5499–5512. [PubMed: 21421562]
24. Smith JD, et al. A method for high-throughput production of sequence-verified DNA libraries and strain collections. *Mol Syst Biol.* 2017; 13:913. [PubMed: 28193641]
25. Savic N, et al. Covalent linkage of the DNA repair template to the CRISPR/Cas9 complex enhances homology-directed repair. *bioRxiv.* 2017; 218149doi: 10.1101/218149
26. Ma M, et al. Efficient generation of mice carrying homozygous double-floxp alleles using the Cas9-Avidin/Biotin-donor DNA system. *Cell Res.* 2017; 27:578–581. [PubMed: 28266543]
27. Gu B, Posfai E, Rossant J. Efficient generation of targeted large insertions in mouse embryos using 2C-HR-CRISPR. *bioRxiv.* 2017; 204339doi: 10.1101/204339
28. Polo SE, Jackson SP. Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. *Genes Dev.* 2011; 25:409–433. [PubMed: 21363960]
29. Chen JS, et al. Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature.* 2017; 550:407–410. [PubMed: 28931002]
30. Kleinstiver BP, et al. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature.* 2016; 529:490–495. [PubMed: 26735016]
31. Slaymaker IM, et al. Rationally engineered Cas9 nucleases with improved specificity. *Science.* 2016; 351:84–88. [PubMed: 26628643]
32. Hu JH, et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature.* 2018; doi: 10.1038/nature26155
33. Ryan OW, et al. Selection of chromosomal DNA libraries using a multiplex CRISPR system. *eLife.* 2014; 3

34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990; 215:403–410. [PubMed: 2231712]
35. Gietz RD, Schiestl RH. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat Protoc.* 2007; 2:31–34. [PubMed: 17401334]
36. SICtools. [Accessed: 17th November 2017] Bioconductor. Available at: <http://bioconductor.org/packages/SICtools/>
37. Sha B, Phillips SE, Bankaitis VA, Luo M. Crystal structure of the *Saccharomyces cerevisiae* phosphatidylinositol-transfer protein. *Nature.* 1998; 391:506–510. [PubMed: 9461221]
38. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. *J Mol Biol.* 1997; 267:727–748. [PubMed: 9126849]
39. [Accessed: 12th February 2018] R: The R Project for Statistical Computing. Available at: <https://www.r-project.org/>
40. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma Oxf Engl.* 2010; 26:139–140.
41. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* 2012; 40:4288–4297. [PubMed: 22287627]
42. ggplot2 - Elegant Graphics for Data Analysis. Hadley Wickham | Springer; Available at: <https://www.springer.com/us/book/9780387981413> [Accessed: 10th February 2018]
43. Hunter JD. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng.* 2007; 9:90–95.
44. Waskom, Michael, et al. seaborn: v0.7.1. Jun, 2016. (Zenodo, 2016)

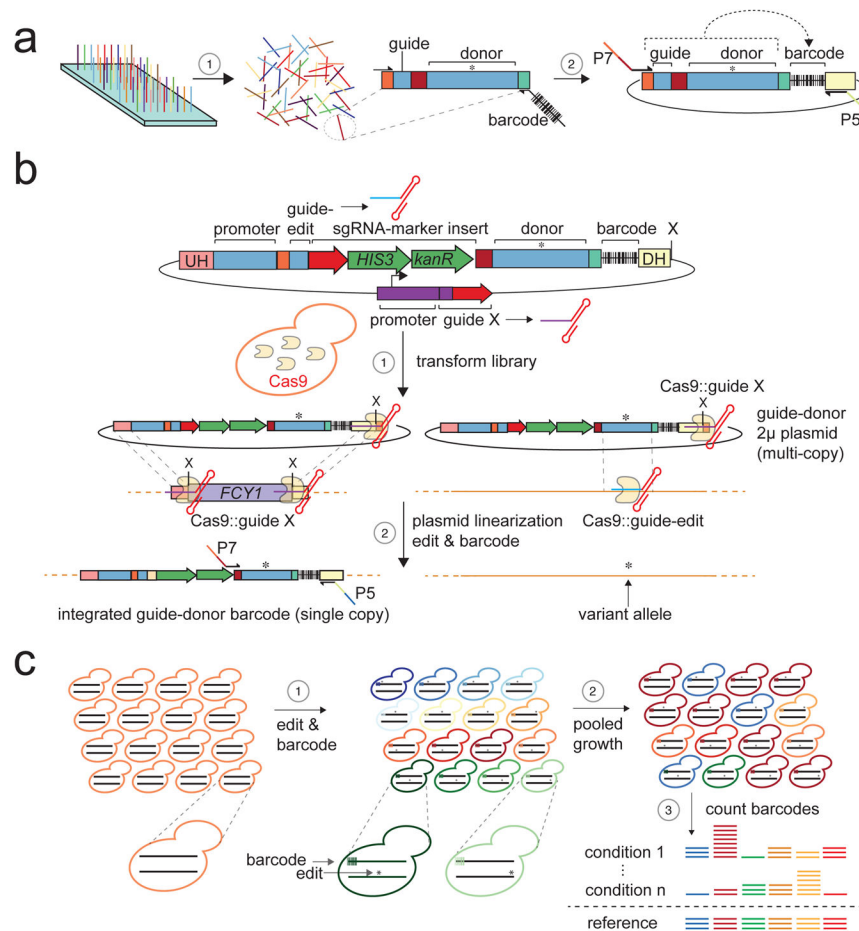


Figure 1. The MAGESTIC pipeline for multiplexed precision genome editing

(a) Linking guide-donors to short DNA barcodes. (1) A complex pool of array-synthesized oligonucleotides encoding guide-donors is amplified and cloned to generate the step 1 library (see Methods). The reverse primer introduces a semi-random 31-mer barcode into each ligation product, and NGS enables sequence validation and computational mapping of each guide-donor sequence in the step 1 library to a unique barcode. **(b)** Insertion of the Cas9 structural guide component plus yeast (*HIS3*) and bacterial (*kanR*) selection markers in between the guide and donor. (1) This final step 2 library is transformed into yeast cells such that the vast majority of transformants uptake a single plasmid which accumulates to high-copy number. Each cell harbors a barcode integration locus with a counter-selectable marker (*FCY1*). Guide-donor plasmids harbor a second guide expression unit (guide X) to promote barcode integration, as guide X cleavage sites flank *FCY1*. Cas9 and guide expression results in simultaneous cleavage of the guide-donor plasmid at a guide X site adjacent to the downstream homology (DH), target site editing (right), and genomic integration of the guide-marker-donor-barcode cassette (left). **(c)** Library-scale genome editing and competitive growth phenotyping. (1) The guide-donor plasmids allow editing throughout the genome, while the barcode integration site is constant. (2) Pooled growth in different conditions results in enrichment or depletion of variants that affect fitness. (3) Variant fold-

changes are calculated based on barcode sequencing counts in treated vs. untreated conditions.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

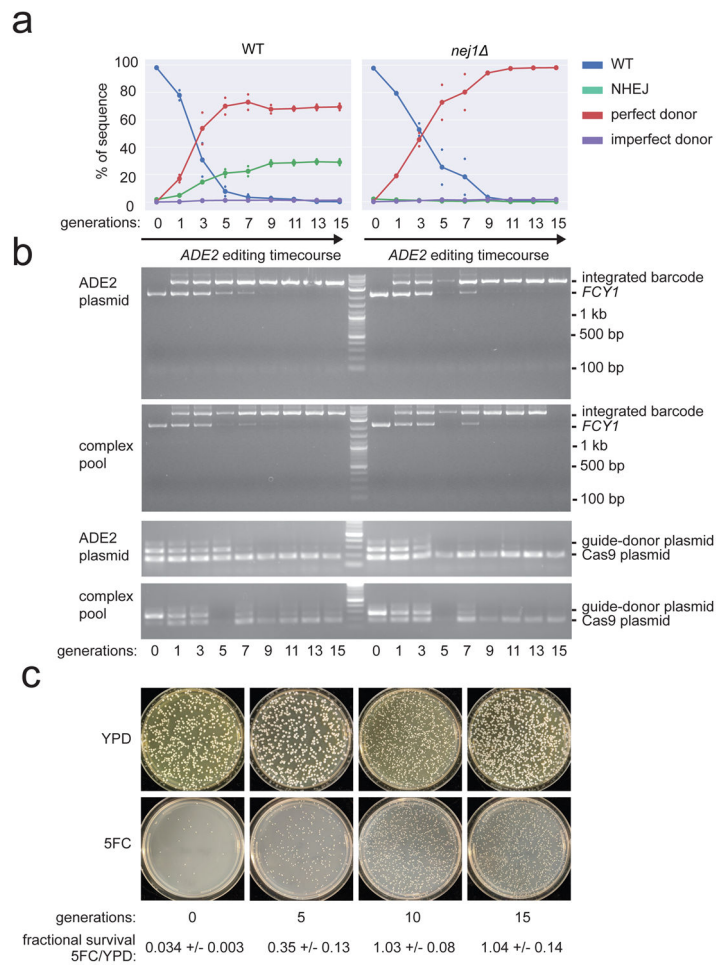


Figure 2. Simultaneous genome editing, guide-donor barcode integration, and plasmid self-destruction

(a) WT and *nej1* were transformed with GAL-Cas9 and a guide-donor cassette to introduce a premature termination codon (PTC) in the *ADE2* gene. Cas9 expression was induced by galactose and aliquots were harvested at the indicated generations. The *ADE2* locus was analyzed by NGS and the fractions of WT sequence, NHEJ indels, and donor DNA-directed editing (either perfect or imperfect repair) were calculated (see Methods). The line graph shows the mean percentages at each generation from duplicate experiments. (b) Integration of the guide-donor barcode was assayed by amplification targeting the chromosomal barcode locus for the single *ADE2* guide-donor plasmid (top) as well as a complex pool of >100,000 barcoded guide-donor plasmids (bottom). The uncropped gel image indicates an absence of detectable NHEJ indel events at the barcode locus. Self-destruction of the guide-donor plasmids was assessed by a three-primer PCR, with a common forward primer and either a guide-donor plasmid-specific primer (top band) or a Cas9-plasmid specific primer (bottom band). (c) Cultures at the indicated generations of galactose induction were plated in quadruplicate at a density of ~1000 cells per plate on rich medium (YPD) and *FCY1* counter-selectable medium (5-FC). The fraction of surviving colonies on plates are shown. All experiments were repeated with three biological replicates starting from independent transformations of the guide-donor plasmids.

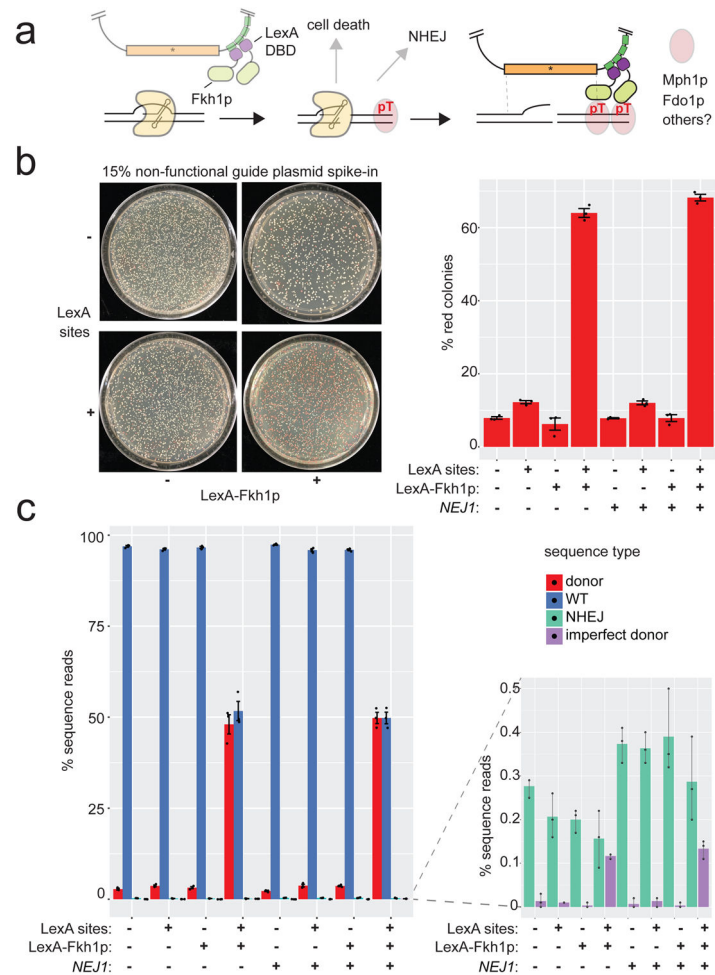


Figure 3. Active recruitment of donor DNA to Cas9-induced dsDNA breaks increases HR efficiency

(a) A protein fusion of Fkh1p to the LexA DNA-binding domain (LexA-Fkh1p) enables recruitment of donor DNA directly to dsDNA breaks (DSBs). DSBs result in the accumulation of proteins phosphorylated on specific threonine residues (pT) near the site of the break. The interaction between Fkh1p and various pT-containing proteins (including Mph1p, Fdo1p, and additional unidentified proteins) recruits LexA-Fkh1p to DSBs, which in turn recruits donor DNA via LexA binding sites the plasmid. (b) *ADE2*-guide donor plasmids with (bottom) or without (top) LexA sites were mixed with a non-functional *ADE2* guide-donor plasmid at a ratio of 17:3, and transformed into a strain pre-expressing TEF1-Cas9 with (right) or without (left) LexA-Fkh1p. Red colonies indicate cells that received a functional *ADE2* guide-donor and survived editing, while white colonies represent cells that received the non-functional *ADE2* guide. The bar chart depicts the mean percentage of red colonies (y-axis) determined by counting 3 plates per condition (x-axis). Error bars represent the standard deviation. (c) The *ADE2* locus was analyzed as in Fig. 2. Because *ade2* is a detrimental mutation, *ade2* null colonies are smaller and thus contribute slightly less sequence reads per colony relative to white colonies. The stacked bar chart (left) indicates

that >99.5% of the sequence is WT or perfect donor repair. The inset bar chart (right) shows the remaining <0.5% of editing events. (see Methods and Supplementary Table 1).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

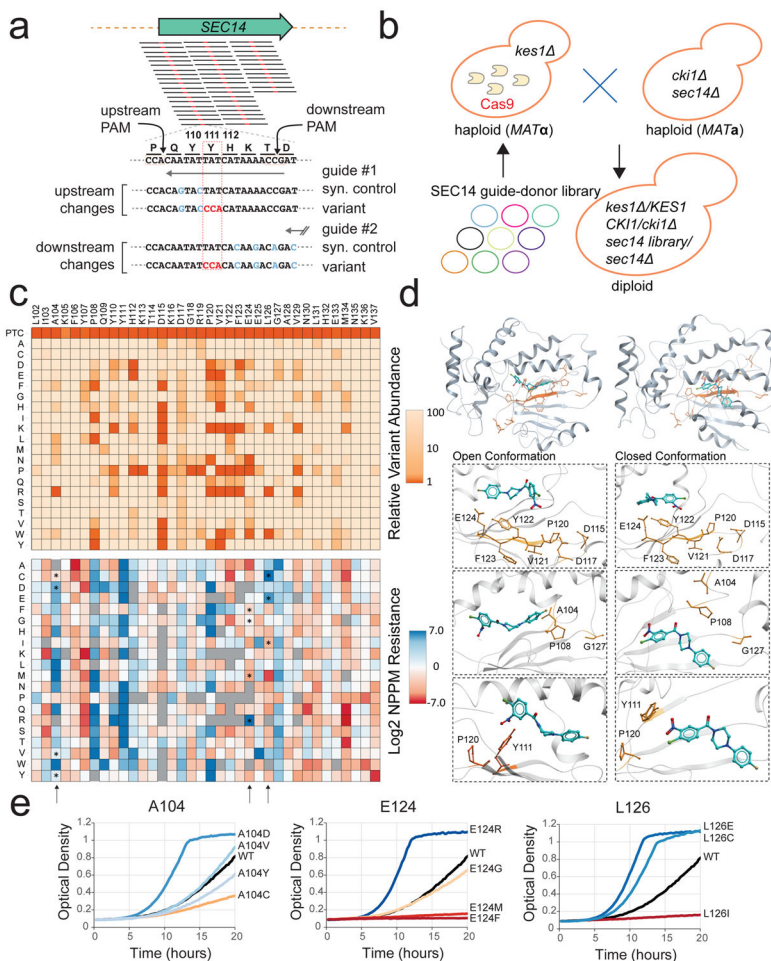


Figure 4. Saturation mutagenesis of an essential eukaryotic gene and structure-function mapping of drug resistance

(a) Synonymous codon spreading strategy for complete saturation mutagenesis of ORFs (see Supplementary Figure 3). (b) Suppressor strategy for assaying function of Sec14p mutants (see Methods). (c) Heat-maps depicting the fitness cost of each mutation in *SEC14* on protein function (top) and resistance to the Sec14 small-molecule inhibitor NPPM (bottom). The normalized relative abundance of each variant in the diploid library was assessed by sequencing (top), and read counts following ~12 generations of growth in 8μM NPPM versus DMSO control were used to calculate relative resistance. Rows indicate specific amino acid mutations, and columns indicate amino acid position in Sec14. Color intensity indicates the degree to which each amino acid mutation negatively impacted Sec14 function (orange, top panel), and increased (blue) or decreased (red) Sec14 resistance to NPPM (bottom panel). The light grey blocks indicate variants with insufficient read counts (see Methods). The screen was conducted in biological replicates. (d) The Sec14 α-carbon backbone (grey) and the mutated window (orange), with the NPPM modeled into both the open (left) and closed (right) conformers of Sec14. Side chains critical for protein function (top panel) and NPPM resistance (middle and bottom panels) are highlighted relative to the predicted binding position for NPPM. (e) The indicated mutants were independently reconstructed and grown with 8 μM NPPM. OD600 was followed over 20 hours of growth,

with WT growth plotted as reference (black line). Growth assays were conducted in biological triplicate with produced nearly identical results (Supplementary Table 2).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

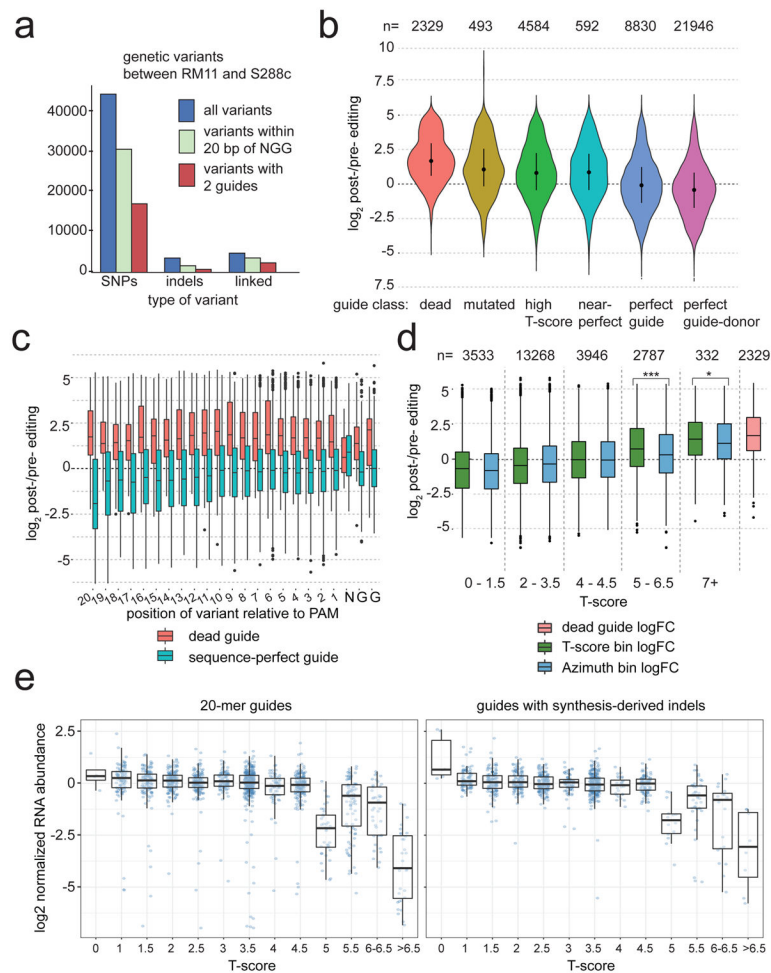


Figure 5. Global profiles of guide efficacy and mismatch tolerance for engineering of natural variants

(a) Number of individual genetic variants between RM11 and S288c (b) Log₂-fold change (logFC) of barcodes in each guide class post-editing vs. pre-editing (dead guides: indels within 15 bp from PAM or ≥ 2 mismatches within 18 bp; mutated guides: 1 mismatch within 18 bp or indel from 16–18 bp; near-perfect: mismatches at positions 19 or 20). High T-score is defined as ≥ 5 (see Methods). (c) Dead and perfect guide-donor logFC by distance of the variant allele from the PAM. Variants shown in the N of the NGG consisted solely of indels or linked variants and harbor additional disruption positions upstream or downstream. (d) Boxplots of logFC by T-score bin with Azimuth score bins of the same size (one-sided Wilcoxon p-values: * = 0.03182, *** = 8.12E-14). (e) Normalized RNA abundances as a function of T-score for sequence-perfect guides (left) and guides with synthesis-derived indels (right). RNA levels were determined by targeted RT-PCR of the guide RNAs and PCR of the guide DNA sequences followed by high-throughput sequencing. RNA and DNA levels were analyzed in biological triplicates and similar results were obtained with random hexamers and with a structural guide-specific primer for reverse-transcription. Box and violin plots show median value, and 25th and 75th quantiles. The

number of barcodes (N) analyzed in each group is shown at the top of each plot for panel b and d, and in the Methods, section statistical analysis, for panel c.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript