High-throughput creation and functional profiling of DNA sequence variant libraries using CRISPR–Cas9 in yeast

Xiaoge Guo1,2,12, Alejandro Chavez1–3,11,12, Angela Tung1,12, Yingleong Chan1,2, Christian Kaas1,2,4, Yi Yin5, Ryan Cecchi1, Santiago Lopez Garnier1,2, Eric D Kelsic1,2,6, Max Schubert1,2, James E DiCarlo1,2,6, James J Collins1,7–10 & George M Church1,2,6,7

Construction and characterization of large genetic variant libraries is essential for understanding genome function, but remains challenging. Here, we introduce a Cas9-based approach for generating pools of mutants with defined genetic alterations (deletions, substitutions, and insertions) with an efficiency of 80–100% in yeast, along with methods for tracking their fitness en masse. We demonstrate the utility of our approach by characterizing the DNA helicase SGS1 with small tiling deletion mutants that span the length of the protein and a series of point mutations against highly conserved residues in the protein. In addition, we created a genome-wide library targeting 315 poorly characterized small open reading frames (smORFs, <100 amino acids in length) scattered throughout the yeast genome, and assessed which are vital for growth under various environmental conditions. Our strategy allows fundamental biological questions to be investigated in a high-throughput manner with precision.

Libraries of cells with defined genetic alterations have proven transformative for connecting poorly understood genes to biological pathways and uncovering novel roles for previously characterized genes. However, in eukaryotes these libraries have been difficult to generate, and even in some widely used collections, such as the yeast knockout library, a majority of the members contain undesired secondary mutations1 and suffer from the presence of selection markers2.

In this work, we present a Cas9-based strategy for the simultaneous, seamless creation of hundreds of genetic variants without integrated selection markers in wild-type yeast cells that express the Cas9 protein along with a donor repair template. Our system is built upon CRISPR–Cas9 and its ability to stimulate homology-directed recombination (HDR) repair of a double-stranded break at a given target locus3. Each isogenic mutant is generated by a plasmid containing a single guide RNA (sgRNA) paired with a corresponding donor template that carries a programmed mutation (hereon referred to as the guide+donor strategy) (Fig. 1a). The advantages of our concatenated guide+donor design are threefold; it enables: a) rapid cloning of all library members within one reaction, b) simultaneous delivery of both the guide and the donor in one contiguous unit thus preventing uncoupling that may result in inefficient repair and unproductive repair outcomes, and c) high-throughput molecular phenotyping using next-generation sequencing (NGS) with guide+donor-containing plasmids serving as unique barcodes for tracking edited cells. A similar concept of in cis delivery of guide+donor was recently demonstrated in bacteria4.

In our initial test, we integrated a copy of the cas9 gene into the neutral HO locus and performed individual transformations of 34 guide+donor plasmids (Supplementary Fig. 1). Upon selecting for cells with the guide+donor, however, we found that the number of colonies with the desired genetic alteration was low (0–30%), consistent with earlier attempts at in cis guide+donor delivery in yeast5 (Supplementary Table 1). We sought to increase the percentage of correctly edited cells in order to enable efficient genome-scale measurements via NGS.

To test if linearization of our guide+donor plasmid would increase the efficiency of our system6–10, we introduced our guide+donor substrate as two linear pieces of DNA. The larger DNA fragment contained the guide+donor portion of the plasmid with an internal portion of the selection marker removed. The smaller DNA fragment consisted of the missing segment of the selection marker with ~150 bp of flanking homology such that HDR was required to reconstitute the full circular plasmid (Supplementary Fig. 1a). With the modified approach, we observed a 6- to 14-fold increase in transformation efficiency (Supplementary Fig. 1b) with 80–100% of the

Received 29 September 2017; accepted 18 April 2018; published online 21 May 2018; doi:10.1038/nbt.4147

1Wyss Institute for Biologically Inspired Engineering at Harvard University, Boston, Massachusetts, USA. 2Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. 3Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA. 4Department of Expression Technologies 2, Novo Nordisk A/S, Maaloev, Denmark. 5Department of Genome Sciences, University of Washington, Seattle, Washington, USA. 6Department of Ophthalmology, Columbia University College of Physicians and Surgeons, New York, New York, USA. 7Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. 8Synthetic Biology Center, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. 9Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. 10Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. 11Present address: Department of Pathology and Cell Biology, Columbia University College of Physicians and Surgeons, New York, New York, USA. 12These authors contributed equally to this work. Correspondence should be addressed to A.C. (ac4304@cumc.columbia.edu) or G.M.C. (gchurch@genetics.med.harvard.edu).
transformants containing the desired repair event, which is in stark contrast to the 0–30% proper editing observed with the unmodified method.1,5 (Supplementary Table 1). No programmed edits were observed in the absence of Cas9 (Supplementary Table 1).

To begin characterizing the limitations of our system, we tested a series of vectors designed to introduce either targeted point mutations, short deletions, or sequence replacements within the ADE2 locus. For programmed point mutations, we obtained a genome modification efficiency close to 100% for changes that occurred proximal to the Cas9-generated cut site (Supplementary Fig. 2). In contrast, when the desired mutation was positioned further away from the Cas9 cut site, we noted a decrease in efficiency, with mutations 12–15 bp away, showing rates of editing of ~40%. While longer homology length (Supplementary Fig. 3a) increased the number of colonies obtained per transformation (Supplementary Fig. 3b), it did not substantially improve the proportion of correct edits (Supplementary Table 2). Of the clones that did not have the desired point mutation, the majority had a mutation in the protospacer adjacent motif (PAM), as designated on the provided guide+donor to escape Cas9 cutting. Further characterization of our method for generating programmed deletions revealed that our design allowed efficient removal of up to 61 contiguous bases (>90% of colonies with the desired change) but experienced a sharp decline in efficiency in creating larger deletions (≥212 bp; Supplementary Fig. 4a). Similarly, our strategy enabled efficient replacement of 61 bp of endogenous sequence with up to 60 bp of user-defined sequence (Supplementary Fig. 4b).

Having gained insight into the limitation of our guide+donor strategy, we next sought to determine the generality of our method by targeting three additional loci (SGS1, SRS2, and ARS214) with a series of point mutations, deletions, and sequence replacements. Similarly to our initial results, we obtained a high efficiency of genome modification (90–100%), across all targets and mutation types (Fig. 1b).

To examine the targeting specificity of our Cas9-based platform, we performed whole genome sequencing on three mutant strains (ade2Δ Ade1Δ Ade2Δ, sgs1Δ Ade1Δ Ade2Δ, and sgs1Δ Ade1Δ Ade2Δ) generated via our guide+donor method and observed the expected genomic edits (Supplementary Fig. 5). Upon surveying all the regions in the genome that had up to two mismatches within the N20 guide sequence, we did not find off-target sites. Off-target effects due to Cas9 are known to result in indels. When the N20 matching parameter was further relaxed to N15+PAM, we did not observe any indels indicative of off-target Cas9 effects.

The strong correlation between the presence of a particular guide+donor plasmid and the presence of the desired genetic

Figure 1 Guide+donor genome-editing platform for engineering and phenotypically characterizing programmed mutations in pool. (a) Illustration of guide+donor workflow. Guide+donors targeting different genomic sites of interest are marked by different colors. Each guide+donor structure contains an SNR52 promoter (yellow), an N20 sequence (dark gray), a structural sgTail (not shown), a terminator sequence (circle-backslash symbol), and a donor template with the desired mutations flanked by regions of homology (red). Pool of transformants is subject to reference and test conditions simultaneously, genomic DNA extraction, and next-generation sequencing of the guide+donor amplicons to determine depletion and enrichment of guide+donor targets. (b) Bar graphs depicting editing efficiencies for creating programmed amino acid substitution, deletion, and sequence replacement at three endogenous sites (ARS214, SGS1, and SRS2). Catalytic amino acid substitutions for SGS1 and SRS2 and proportion of correct edits are indicated. (c) Graphical representation of guide+donor-generated ARS214 (gray) and SGS1 (red) variants. Asterisk, dotted box, and solid dash denote amino acid substitution, deletion, and replacement of an amino acid stretch with a linker sequence, respectively. Figures not drawn to scale. (d) Dot plot of HU response of a guide+donor library of ARS214 and SGS1 mutants. X- and y-axes correspond to programmed edits encoded in the guide+donor constructs and log_2 fold change, respectively. Two independent yeast library transformations were performed. (e) Dot plot of sensitivity of ARS214 and SGS1 mutants in mms4Δ genetic background. Genetic modifications and log_2 fold-change are exhibited on x and y axes, respectively, wo independent library transformations were performed.
alteration should allow us to infer the fitness effects of these modifications by sequencing the abundance of different guide+donor pairs within a mixed pool. To test this hypothesis, we built a small library containing a mixture of guide+donor plasmids designed to modify either the non-essential ARS214 locus or the DNA damage repair helicase SGS1 (Fig. 1c). Cells obtained from the pooled transformation were grown in media with or without the genotoxic agent hydroxyurea (HU) and the abundance of various guide+donor plasmids within the population was determined by NGS. As expected, we observed a marked depletion of guide+donor pairs encoding modifications that disrupted the ATPase domain of SGS1, which is known to play a critical role in SGS1 function (Fig. 1d)\(^{11-15}\), whereas mutating the less essential C-terminus\(^{11}\) led to less depletion (Supplementary Fig. 6). When we introduced synonymous changes within the ATPase domain or C-terminus of SGS1 we did not observe depletion, suggesting that the effects were not due to non-specific disruption of the SGS1 locus by Cas9. Furthermore, when each of the generated strains was tested individually, the results correlated well with our pooled analysis, lending additional support for the validity of our method (Supplementary Fig. 7).

In addition to exposing the mutant library to environmental perturbations, we also asked whether our system could be used to observe gene–gene interactions by transforming our small library into cells defective in the structural endonuclease Mms4. In an mms4Δ genetic background, all SGS1 mutants in the library exhibited a fivefold depletion, consistent with known synthetic sickness between SGS1 and MMS4 (Fig. 1e)\(^{16,17}\).

We subsequently applied our method to perform systematic characterization of a single protein and targeted SGS1, a gene that encodes the yeast homolog of the human DNA helicase BLM with known roles in mitotic stability, cancer, and aging\(^{18}\). To map the critical domains within Sgs1 that provide cellular resistance to the genotoxic stressor HU, we designed a set of guide+donor constructs that generated 20 amino acid deletions with 5 amino acid sliding windows across the majority of the SGS1 gene. Among the regions showing strongest depletion within edited cells were guide+donors deleting amino acid stretches 1–85, 686–1,090, and 1,116–1,225, which correspond to the Sgs1-Top3-binding domain, Sgs1-helicase, and RQC domains, respectively (Student’s two-tailed t-test, \(P < 0.0001\); Fig. 2a and Supplementary Data 1)\(^{19-22}\). These results are consistent with the known mechanism by which Sgs1 functions through the recruitment of accessory proteins (through N-terminal residues)\(^{12,14,15,21,23-27}\) and by resolution of DNA structural intermediates via its helicase and RecQ domains\(^{12,28}\). We performed biological replicates of our library experiments to assess reproducibility and observed a correlation of 0.86 between the \(\log_2\) fold-change (FC) observed in the two independent yeast transformations (Fig. 2b). Furthermore, we performed individual phenotypic validation of seven hits from the library screen via spot assay, and observed similar results (Fig. 2c).

Next, we created a series of precise point mutations within Sgs1. Toward this goal, we selected a set of nine evolutionarily conserved amino acid residues within the Sgs1 helicase domain and attempted to change them to all other possible amino acids using our guide+donor strategy. This library was exposed to increasing concentrations of HU to assay for mutant drug sensitivity. Despite targeting highly conserved residues within Sgs1, all but one tolerated alanine substitution without causing an obvious loss in resistance to our highest concentration of HU at 40 mM (Fig. 3a and Supplementary Data 2). In the case where activity was lost, alanine was used to replace the essential helicase catalytic residue K706. We observed a strong correlation between independent biological replicates (Fig. 3b–e).

Selecting one representative pair of biological replicates (40 mM), we observed a correlation of 0.88 between the first and second biological replicate (Fig. 3e). We individually validated six variant hits from the library screen and observed concordant results (Supplementary Fig. 8). Overall, we observed, as expected, that amino acid substitutions of similar charge and size were well-tolerated while those with the opposite properties were more detrimental to Sgs1 function.

To determine the capacity of our method to perform targeted editing across the entire yeast genome, we designed and built a guide+donor library for generating small deletions around the initiating ATG for a set of 307 randomly chosen canonical ORFs (including both essential and non-essential genes), along with 315 poorly characterized smORFs. Unlike canonical ORFs, smORFs remain largely ignored and are often missing in modern genome annotations due to their size, low conservation scores, and lack of similarity to known proteins and protein domains.

Using our genome-scale deletion library, we first performed an essentiality screen. We observed strong depletion (~8- to 100-fold) for all targeted essential ORFs (two-tailed t-test, \(P < 0.0001\)) compared to about a threefold depletion for nearly all nonessential ORFs (two-tailed t-test, \(P = 0.01\)), thus highlighting the specificity and sensitivity of our method (Fig. 4a and Supplementary Data 3). Out of the smORFs that were examined, 19 smORFs showed similar levels of depletion as our essential controls (two-tailed Z-test, \(P < 0.001\), in line with previous results\(^{29}\). When we repeated our screen, we observed a correlation of 0.71 between the two independent biological replicates (Supplementary Fig. 9).

Although a number of our smORF library members were located in close proximity to essential ORFs (in some cases within 132 bp), our screen did not identify any of them as essential, emphasizing the specificity of our targeting method. To further demonstrate the ability of our guide+donor strategy to characterize a large number of proteins in parallel, we subjected our smORF mutant library to a series of environmental stressors including growth: at 37 °C (Fig. 4b), in the presence of HU (Fig. 4c), or with the antifungal drug fluconazole (Fig. 4d). For each of our screens, we identified nearly all of the previously known smORFs with tolerance toward each of the tested conditions, along with uncovering previously unreported roles for a large number of additional smORFs\(^{30}\). We individually validated 13 of the hits from our library screens and observed phenotypes in agreement with the screen results (Supplementary Fig. 10).

Of the 315 smORFs examined, 68 were found to play a role in cellular fitness under test conditions. This is in contrast to conventional ORFs for which 104 of 307 tested ORFs were found to be involved in growth under the same environmental conditions (Chi-squared test, \(P < 0.0001\)). Next, we examined features (including amino acid size, gene expression level, secondary structure formation, and evolutionary conservation) that could be shared by the smORFs or the ORFs exhibiting biological activity. Although smORFs show a range of sizes across the yeast genome (smallest smORF hit was 28 amino acids), we found that longer smORFs with elevated levels of RNA expression exhibited a trend of being more likely to come up as hits in our screen (Supplementary Table 3). Notably, ORFs showed no such correlation with regard to length, but maintained a similar trend with respect to expression (Supplementary Table 4). Moreover, we did not observe any difference in the prevalence of structural elements (e.g., alpha-helices and beta-sheets) within smORF hits as compared to non-hits. We did, however, observe an increased propensity for beta-sheets and a decrease in unstructured loops when smORFs as a whole were compared to the set of ORFs that was also examined in our screens (Supplementary Table 5). Finally, a large difference in the rate of gene
conservation was found with 32 of the 68 smORF hits being conserved in humans as compared to only 43 of the 247 smORFs that showed no effect upon the examined conditions (Chi-squared test, \( P < 0.0001 \) (Supplementary Table 6)).

Here, we present a high-throughput method for the rapid generation and phenotypic characterization of hundreds of mutants and illustrate its potential in domain/residue mapping and functional interrogation of nearly any user-defined genomic target by introducing deletions, amino acid substitutions, and sequence replacements. This enables the creation of specific user-defined loss-of-function, gain-of-function, and altered regulation mutants en masse.

By editing the locus within its native context without the need for exogenous markers, we avoided artifacts from using surrogate reporter systems and false-positive and false-negative results due to selection-marker-driven positional effects (Supplementary Table 7). The high library editing efficiency of our system (85–95%) (Supplementary Table 7) allows users to read the guide+donor sequence on the plasmid delivered to each cell and use the sequence to identify the cell’s genotype.

Ultimately, this feature enables the fitness of hundreds, potentially thousands, of mutants to be tracked by sequencing the abundance of each guide+donor sequence within a population. While our method employs a similar gap-repair mechanism as reported by Horwitz et al.\textsuperscript{32}, our design is unique in that each guide is concatenated to a corresponding donor repair template, enabling simultaneous delivery of guide+donor.

Our tiling deletion experiment on SGS1 demonstrated our technology’s ability to rapidly home in on the critical domains required for protein function. A similar CRISPR-based protein perturbation concept to identify critical functional domains in mammalian cells\textsuperscript{33,34} and in yeast\textsuperscript{35} was reported previously. Of note, the underlying mechanisms of functional perturbation between these aforementioned two systems and our guide+donor platform are different in that the former ones rely on unpredictable CRISPR-induced indel and random transposase-induced insertion mutagenesis, respectively, while in our method the variants are created through programmed genetic alterations.

![Figure 2](https://example.com/figure2.png)

**Figure 2** Guide+donor library of sgs1 mutants in response to HU. (a) Sgs1 tiling deletion screen. Scatter plot showing average log\(_2\) fold-change in abundance of guide+donor members programmed to generate sgs1 tiling deletion mutants across the entire SGS1 gene in response to HU \( (n = 2 \) independent yeast library transformations). Guides paired with corresponding donor sequences to generate programmed deletions are indicated in blue. Non-targeting control guides paired with sequence that lack homology regions to qualify as donors are used as controls and are shown in orange. x and y axes denote the amino acid window along the protein and average log\(_2\) fold-depletion, respectively. Schematic representation of relevant domains and motifs in Sgs1 is shown. Figures not drawn to scale. (b) Replicate analysis of log\(_2\) fold-changes between two independent yeast library transformations. Pearson correlation coefficient is indicated. (c) Phenotypic validation of selected sensitive and non-sensitive sgs1 truncation mutants from the HU library screen in a.
Figure 3 Guide+donor library of amino acid substitutions of selected conserved residues in SGS1 in response to various concentrations of HU. (a) Sgs1 amino acid residue substitution screen. Scatter plots showing average log_2 fold-change in abundance of guide+donor members programmed to generate precise point mutations within Sgs1 in response to HU (n = 2 independent yeast library transformations). Concentrations of HU are represented by different colors and described in the legend. Selected conserved residues and average log_2 fold-depletion are displayed on the x and y axes, respectively. Each subplot shows the corresponding amino acid by which each conserved residue was replaced. (b–e) Replicate analyses showing Pearson correlation of log_2 fold changes between two independent yeast transformations under various drug concentrations.
Deep mutational scanning (DMS) methods provided a framework for generating point mutations in a single protein of interest and functionally annotating a large fraction of these amino acid substitutions. However, these methods are only meant to interrogate a single gene at a time, which hinders the scale of functional genomics experiments one can perform. In addition, many deep mutational scanning methods are carried out on a plasmid, thus taking the examined protein variant out of its native context. Although our amino acid substitution library was not as exhaustive in its targeting scope as DMS, we were able to target hundreds of genes at a time and perform all of our genetic alterations within the native genomic locus. Previous work by Kastenmayer et al. used labor-intensive conventional techniques to make specific gene deletions of 140 smORF mutants. In contrast, we demonstrated the ease of our guide+donor method in rapidly covering over ~79% of the 299 putative smORFs within the yeast genome, including many that had previously been neglected. Given the degree of conservation between yeast and human genomes and the conservation between several smORFs and higher eukaryotes, it will be interesting to see if the smORFs identified in our work with roles in stress tolerance have similar functions in humans.

Our method employs the commonly used Streptococcus pyogenes Cas9 (SpCas9), which limits the potential target sites because of its PAM-specific requirement. Using Cas9 variants recognizing alternative PAMs could greatly broaden the range of sequences that can be modified by our approach.

Although we have focused on the usage of our technology for high-throughput characterization of coding elements, we envision a broad range of additional applications, such as, directed evolution, metabolic engineering, and functional interrogation of non-coding elements. Moreover, given that most clinically relevant mutations are point mutations and given the high degree of gene conservation between yeast and humans, our guide+donor editing platform provides an easy way to engineer and test the effects of hundreds of currently uncharacterized single-nucleotide polymorphisms that exist within human populations via their nearest yeast ortholog.

**METHODS**

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

**ACKNOWLEDGMENTS**

G.M.C. was supported by NIH grants R01 HG008525 and P50 HG005550. A.C. was funded by the National Cancer Institute grant no. ST32CA09216-34. J.J.C. was funded by the Defense Threat Reduction Agency grant HDTRA1-14-1-0006, the Paul G. Allen Frontiers Group. Y.Y. was supported by the Damon Runyon Research Foundation grant DRG-2248-16.

**AUTHOR CONTRIBUTIONS**

X.G. and A.C. conceived the idea, led the study, and designed all experiments. A.C. and R.C. with input from J.E.D. demonstrated the initial feasibility of the guide+donor approach. X.G. performed majority of the experiments, including the oligo library design, library construction and analysis, with significant technical contribution from A.T. Y.Y. provided expertise in statistical analysis. Y.Y. performed the whole genome sequencing experiment for off-target analysis.
Functional profiling of the Engineered CRISPR-Cas9 nucleases with altered PAM. Sgs1 helicase defines an essential function for Damage tolerance protein Mus81 associates with the FHA1 Functional genomics of genes with small open reading frames. Efficient multiplexed integration of synergistic alleles...
ONLINE METHODS

Yeast strains and growth conditions. All strains were derived from YAC2370 (BY4741 derivative; MATa his3A1 leu2A met15A ura3A). YAC2563 was constructed by one-step integration of a PmlI-linearized plasmid carrying human-codon-optimized Cas9 with expression under the control of the NOP1 promoter along with a linked NatMX drug selection marker (ACG6218) into the HO locus. MMS4 was deleted in YAC2563 background by one-step gene replacement using PCR-generated deletion cassettes (mms4Δ·KanMX).

Cells were grown non-selectively in YPAD (1% Bacto-yeast extract, 2% Bacto peptone, 2% dextrose; 1.5% agar for plates) supplemented with 500 µg/ml adenine hemisulfate. Ura+ colonies were selected on synthetic complete (SC) medium deficient in uracil (SC-Ura). All growth was at 30 °C. For the experiments with SGS1 mutants, hydroxyurea (Sigma–Aldrich) was added to final concentrations of 5 mM, 10 mM, 20 mM, and 40 mM. For the smORF library drug conditions, flucanazole (Sigma–Aldrich) and HU (Sigma–Aldrich) were added to final concentrations of 25 µg/ml and 100 mM, respectively.

Plasmids. Guide+donor plasmids were built in the yeast pRS426 2 µm back- bone containing the URA3 selection marker. The guide RNA expression cassette contained SNP52 promoter, guide RNA sequence, chimeric single-guide RNA structural tail (sgtail), and SUP4 terminator. The donor sequence carrying the desired modification was placed immediately downstream of the terminator sequence. Individual guide+donor fragments were generated from three overlapping PCR fragments using 90-mer oligos from IDT designed to create the guide sequence and its corresponding donor sequence. The ends of the stitched PCR amplicon were designed such that they contained overlapping regions for Gibson assembly. These fragments were then assembled in combination with the plasmid backbone that was digested with NgobMV and Nhel to prepare it to accept the incoming guide+donor sequence. For library cloning described below, the plasmid backbone was further modified to remove BsmBI and SapI sites.

Guide+donor library design. Custom Python scripts were used to design the libraries. Oligos were synthesized by CustomArray Inc. For the SGS1 tilting deletion library with a sliding window of 15 bp, we generated donor sequences with 80 bp total homology flanking each 60-bp deletion region, then coupled a 20-bp guide RNA that was present in each deletion region closest to the mid- dle of the section being removed. For the Sgs1 amino acid library, we targeted the conserved residues previously reported by Kusano et al. (1999)43 and also included the known catalytic residue lysine 706 (K706) as a positive control. The N20 was positioned closest to the target residue and 80-bp donors were designed to change the conserved target residue to every other amino acid. Finally, the smORF deletion library was designed to delete 60 bp from the 5’ terminus of each target, including the initiating ATG when possible. SapI sites were added between the guide and the donor sequence that was synthesized by CustomArray to enable downstream cloning of the sgtail and an RNA polIII terminator between these two elements. Finally, all synthesized oligos had BsmBI sites added to each end to enable the first stage of cloning in which the oligo library members were inserted into the pRS426 backbone. Library members containing restriction sites including BsmBI, SapI, NcoI, and Stul were excluded from the sequence file and were not synthesized.

Cloning of the library. The CustomArray-synthesized oligo library was diluted to 1 ng/µl and 1 µl of the library was amplified with Kapa SYBR FAST qPCR Kit Master Mix (Kapa Biosystems) using unique primer pairs specific to each desired library (e.g., SGS1 tilting deletion, smORF library, etc.). Primers used for oligo library amplification were further modified to contain the necessary overlaps to enable the library to be inserted into our vector backbone via Golden Gate cloning. The PCR products were run on a gel to confirm amplicons were of the expected length. After PCR purification (Zymo Research), the amplicon was cloned into the BsmBI-containing vector (XG128) using a standard Golden Gate protocol with BsmBI (NEB R0580S) and T4 ligase (NEB M0202S) then electrooporated into 5-alpha electrocompetent Escherichia coli cells (NEB C2989). This ensuing library now contained the guide and donor sequences adjacent to the SNP52 promoter but was still missing the sgtail and an RNA polIII terminator. To clone in the additional functional components between the guide and donor, we amplified and cloned in the sgtail and terminator sequences following the same Golden Gate cloning method as described above, but this time using SapI (NEB R0569S) and T4 ligase. The resulting Golden Gate reactions were then PCR-purified and electroporated into 5-alpha electrocompetent E. coli cells to create a final guide+donor library.

Transformation into yeast. Prior to transformation into yeast, each guide+donor library was double-digested with NcoI (NEB R0193T) and Stul (NEB R0187L), resulting in a linearized vector with a gap within the URA3 selection marker. Linearized DNA containing the majority of the vector backbone, but lacking a portion of the URA3 selection marker, was then gel extracted and purified (Zymo Research). To enable the reconstruction of the guide+donor vector within yeast via homologous recombination, a second linear fragment was generated by PCR using primers that annealed to regions flanking the NcoI and Stul restriction sites, creating a PCR fragment with >100 bp of overlap homology to the region removed from the guide+donor backbone. Digested DNA and PCR amplicons (1 µg each per transformation) were co-transformed into yeast using standard lithium acetate transformation protocol with the addition of dimethyl sulfoxide (DMSO, 10% final concentration) before heat shock and grown on SC-Ura plates for 3 d to obtain Ura+ colonies.

For our initial library pilot experiments (Figs. 1b,d,e and Supplementary Figs. 6b,c), 500 ng of each indicated guide+donor plasmid was pooled and double-digested with NcoI and Stul. 1 µg of the linearized plasmid mix was co-transformed with 1 µg of Ura3 PCR fragment (as described above) into Cas9-expressing wild type and Mms4-inactivated strains in parallel and selected on SC-Ura. Ura+ colonies were scraped off plates after 3 d. For HU sensitivity screen, cells were further diluted 1:100 in liquid media that contained either no HU or 40 mM HU, and grown for 2 d. Cells were collected and genomic DNA was extracted for NGS. Two rounds of independent yeast transformation were performed.

For the HU condition test of the SGS1 mutant libraries, each library was first transformed into no-Cas9 and Cas9-expressing cells in parallel using the yeast transformation procedures as described above and selected on SC-Ura. After 3 d, colonies were scraped off the plates, diluted 1:100 in liquid media that contained either no HU or 40 mM HU, and grown for 2 d. Cells were then collected and genomic DNA was extracted for NGS. Experiments were done in duplicate.

For the essentiality/non-essentiality test of smORF library, the library was transformed into no-Cas9-expressing cells and Cas9-expressing cells in parallel. Colonies were scraped and diluted 1:100 in liquid media and grown for 2 d. In addition, transformants from the Cas9-expressing cells were also grown in liquid media containing either 100 mM HU, 25 µg/ml flucanazole, or subject to 37 °C for 2 d. Subsequently, cells were collected, genomic DNA extracted, and NGS was performed. All experiments were done in duplicate.

Guide+donor library preparation and sequencing. Genomic DNA was isolated from each yeast sample. Two rounds of PCR were performed using Q5 Hot Start High-Fidelity polymerase (New England BioLabs). The first round amplified each guide+donor with forward (CCTTTCCCTACAGCGCTC TTCCGATCTNNNNNAGTGAAAGATATATGATC) and reverse primers (GGAGTTCCAGCAGTTGCTCCTTCGATCTGGAATGATTGATCGACTG) hybridizing to common flanking regions. Subsequently, standard Illumina TruSeq and/or Nextera barcodes were attached through a second round of PCR amplification. Gel purification was performed on all amplicons to confirm the amplicon size and quality before extracting and purifying the sample using the QIAquick gel extraction kit (Qiagen). DNA libraries for NGS were quantified using the Kapa Library Quantification Kit (Kapa Biosystems). Samples were pooled in equimolar amounts. The final library was prepared using standard MiSeq Reagent Kit v2 (2 × 150 bp) protocol with 12 µM diluted DNA libraries with 15–25% PhiX spiked into the mixture and run on an Illumina MiSeq or NextSeq 500 Systems, respectively.

Preprocessing of library sequences and count generation. Guide RNA and donor sequences were extracted from R1 and R2 reads, respectively, and matched to reference library members containing each guide+donor pair using a custom Python script. Sequences that did not match any of the library members...
were removed from the analysis. Only sequences that contained the perfectly matched N20, sgtagl (GTTCATAGCCGTAAGAATTCGAATCGGTGAAACGTAACATGAGTTCGTTATCAACTGTGAAAGATGGGACCCAGGTGTTGTTGTTTGTGGTTATGCTT) and donor sequences were included in count generation. We first sequenced the plasmid libraries to determine the distribution of sequences. Reads that were severely underrepresented, that is, less than 30 reads mapped to the guide+donor, were removed from further analysis.

Data analysis and fitness calculation. For all the conditions, the mapped reads were compared against the corresponding control experiment. The control experiment for each SG51 library (tiling deletion and amino acid substitutions) was the experiment performed in the absence of HU. A fold-change (FC) for each guide+donor is calculated as follows:

$$FC_i = \frac{test_i}{test_{total}} \times \frac{control_{total}}{control_i}$$

where test and control are the number of reads that mapped to guide+donor i in all the test conditions and control, respectively. The test total represents the total number of reads in the test conditions and control total is the number of reads in the control. If the guide+donor is enriched in the test condition, FC would be >1. If the guide+donor is depleted in the tested condition, FC would be <1. The average log2FC values of the duplicates and P-values (two-tailed Z-test) corresponding to each tested guide+donor for each library are provided in Supplementary Data 1 and 2.

The smORF library was subjected to four screens: essentiality, heat, HU, and fluorocitazole. While the control experiment for the last three test conditions was conducted in the absence of the environmental stress, the control experiment for essentiality screen was performed in a yeast strain lacking Cas9. The same FC calculation described above was carried out for each guide+donor in the smORF library. Supplementary Data 3 lists the average log2FC value and P-value for each tested guide+donor.

Validation of mutants from the three libraries. Individual Ura+ transformants were picked from each library and grown overnight in 96-well plates. DNA extraction was performed, followed by PCR amplification (forward primer TTTGTTAGCGTTATCAACTGTGAAAGATGGGACCCAGGTGTTGTTTGGTGGTTATGCTT and reverse primer GTTCATAGCCGTAAGAATTCGAATCGGTGAAACGTAACATGAGTTCGTTATCAACTGTGAAAGATGGGACCCAGGTGTTGTTTGGTGGTTATGCTT) and sequencing of the guide+donor on the plasmid (TTCGCAGGTTCAAACTCTCCCGCA) to determine programmed edits intended for each transformant. Individual primer pairs specific to the corresponding endogenous site were designed. Each endogenous site was amplified and sequenced with the forward primer to determine if the programmed edits as specified by the donor had successfully occurred.

Phenotypic validation of library hits. To validate the hits exhibiting phenotypic sensitivity and lack of sensitivity in each library screen, we picked two to four sensitive and two to four non-sensitive targets from each screen, constructed the corresponding guide+donor plasmids, and performed similar transformation experiment as described above. Individual transformants were genotyped followed by phenotyping onto the corresponding test conditions to confirm our NGS screening results. For the phenotypic growth assay, cells were grown to log phase. 3 μl of each undiluted and fivefold serially diluted culture were spotted onto SC-URA or SC-URA under tested conditions. All plates were incubated at 30 °C for 48 h and photographed.

Whole genome sequencing to detect off-target effects of guide+donor system.

Sample preparation. A Cas9-expressing parental yeast strain (YAC2563) and three yeast strains (YXG231, YXG232, YXG234) modified by guide+donor plasmids, ADE2Δ61bp, SGS1Δ60bp, and SGS1ΔATG, respectively, were grown overnight in 5 ml YPAD. Genomic DNA was isolated from these cells following a PCR purification (Zymo Research) step to clean up the DNA. For library preparation, we used Nextera (Illumina) to fragment the genome. Roughly 35 ng of genomic DNA was used for each sample, equivalent to 3 million haploid yeast genomes. After the fragmentation reaction (20 μL reaction system, 55 °C 15 min, 70 °C 30 min), fragmented DNA was purified with DNA Clean-Up & Concentrator-5 (Zymo Research) and used as PCR template (NEBNext High-Fidelity 2X PCR Master Mix, NEB, 72 °C 3 min for Tns5 gap filling and end repair, 98 °C 30 s, 4 cycles of 98 °C 10 s, 63 °C 30 s, 72 °C 40 s, and 72 °C 2 min for a final extension) to add sequencing adaptors. Amplified library
was cleaned up with 0.8×x Ampure beads and sequenced with 2 × 150 bp NextSeq500/550 for a total of 28 M paired-end reads.

**Computational analysis.** The quality of the fastq files was first evaluated using the FASTQC tool ([https://www.bioinformatics.babraham.ac.uk/projects/fastqc/](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)) followed by end trimming using FASTX Toolkit ([http://hannonlab.cshl.edu/fastx_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)) to obtain base pairs with a quality score lower than 30 (fastq_quality_trimmer). After quality trimming, an in-house algorithm was used to intersect the read-pairs. Subsequently, BWA (version 0.6.1-r104) was used to align the reads to the S288C genome downloaded from GenBank as assembly GCA_000146045.2. SNPs were detected by SAMtools mpileup and bcftools. A hard filter removing all SNPs/indels below 25% of the median depth was chosen as cutoff. The median depth was deduced using genomeCoverageBed from BEDTools (version 2.16.2) as described by Kaas *et al.*47. Off-target analysis was carried out using Bowtie (version 0.12.7) to search the yeast genome for the guide RNA sequences corresponding to the guide+donor constructs for up to two mismatches. A region of 500 bp surrounding each of the potential off-target sites were manually cross-referenced with the list of detected SNP/indels as previously described48. The expected genomic changes were manually evaluated from the aligned BAM file in Geneious (Biomatter Ltd.).

**Statistical analysis.** Each figure description indicates the number of independent experiments. A two-tailed $Z$-test was used to examine significance of depletion in each library screen in Figures 2 and 3. A two-tailed $Z$-test was applied to examine depletion and enrichment in Figure 4. A Chi-squared test was used to assess statistical differences between groups in Supplementary Tables 1 and 2. A two-tailed $t$-test was to examine statistical significance in Supplementary Tables 3 and 4. Kolmogorov–Smirnov and Chi-squared tests were used to assess the statistical differences between groups in Supplementary Tables 5 and 6, respectively.

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

**Data availability.** All NGS data generated in this study are available under SRA accession numbers SRP140162, SRP140351, SRP140360, SRP140255, SRP140260. Data used for amino acid length, gene expression, and human conservation comparisons are presented on Supplementary Data 4 and summarized in Supplementary Tables 3–6.

**Code availability.** All custom scripts are available upon request.

41. Christianson, T.W., Sikorski, R.S., Dante, M., Shero, J.H. & Hieter, P. Multifunctional yeast high-copy-number shuttle vectors. *Gene* **110**, 119–122 (1992).
42. Kusano, K., Berres, M.E. & Engels, W.R. Evolution of the RECQ family of helicases: a drosophila homolog, Dmblm, is similar to the human bloom syndrome gene. *Genetics* **151**, 1027–1039 (1999).
43. Yao, W. *et al.* The INO80 complex requires the Arp5-les6 subcomplex for chromatin remodeling and metabolic regulation. *Mol. Cell. Biol.* **36**, 979–991 (2016).
44. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
45. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
46. Rost, B., Yachdav, G. & Liu, J. The PredictProtein server. *Nucleic Acids Res.* **32**, W321–6 (2004).
47. Kaas, C.S., Kristensen, C., Betenbaugh, M.J. & Andersen, M.R. Sequencing the CHO DXB11 genome reveals regional variations in genomic stability and haploidy. *BMC Genomics* **16**, 160 (2015).
48. Paix, A. *et al.* Scalable and versatile genome editing using linear DNAs with microhomology to Cas9 Sites in Caenorhabditis elegans. *Genetics* **198**, 1347–1356 (2014).
Life Sciences Reporting Summary

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

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

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study. For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

1. Experimental design

   1. Sample size
      Describe how sample size was determined.
      No statistical methods were used to predetermine sample sizes. Sample size choices were made based on the magnitude and consistency of measurable differences between groups.

   2. Data exclusions
      Describe any data exclusions.
      In our library experiments, sequences that did not match to designed guide+donor library members were not included in the analysis.

   3. Replication
      Describe the measures taken to verify the reproducibility of the experimental findings.
      Yes, as described in figure legends and Methods.

   4. Randomization
      Describe how samples/organisms/participants were allocated into experimental groups.
      No randomization was used for samples

   5. Blinding
      Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
      Investigators were not blinded to sample allocation during data collection

   Note: all in vivo studies must report how sample size was determined and whether blinding and randomization were used.

6. Statistical parameters

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

   □ n/a
   □ Confirmed

   - The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
   - A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
   - A statement indicating how many times each experiment was replicated
   - The statistical test(s) used and whether they are one- or two-sided
   - A description of any assumptions or corrections, such as an adjustment for multiple comparisons
   - Test values indicating whether an effect is present
   - Provide confidence intervals or give results of significance tests (e.g. P values) as exact values whenever appropriate and with effect sizes noted.
   - A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
   - Clearly defined error bars in all relevant figure captions (with explicit mention of central tendency and variation)

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

Nature Biotechnology: doi:10.1038/nbt.4147
Software

Policy information about availability of computer code

7. Software

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

VassarStats (http://vassarstats.net/) and R (version 3.4.2) were used for statistical analyses. Processing of raw reads was performed with the FASTX toolkit (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Python 2.7.10 was used to run custom Python scripts.

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

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a third party.

No unique materials were used.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

We did not employ eukaryotic cells we used yeast strain BY4741 in our experiments.

b. Describe the method of cell line authentication used.

N/A

c. Report whether the cell lines were tested for mycoplasma contamination.

N/A

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

N/A

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

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

No animals were used in our studies.

Policy information about studies involving human research participants

12. Description of human research participants

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

No human research participants were used in our studies.