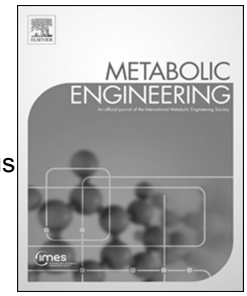
Accepted Manuscript



Validating genome-wide CRISPR-Cas9 function improves screening in the oleaginous yeast *Yarrowia lipolytica*

Cory Schwartz, Jan-Fang Cheng, Robert Evans, Christopher A. Schwartz, James M. Wagner, Scott Anglin, Adam Beitz, Weihua Pan, Stefano Lonardi, Mark Blenner, Hal S. Alper, Yasuo Yoshikuni, Ian Wheeldon

PII: S1096-7176(19)30112-0

DOI: https://doi.org/10.1016/j.ymben.2019.06.007

Reference: YMBEN 1561

To appear in: *Metabolic Engineering*

Received Date: 13 March 2019

Revised Date: 6 June 2019

Accepted Date: 14 June 2019

Please cite this article as: Schwartz, C., Cheng, J.-F., Evans, R., Schwartz, C.A., Wagner, J.M., Anglin, S., Beitz, A., Pan, W., Lonardi, S., Blenner, M., Alper, H.S., Yoshikuni, Y., Wheeldon, I., Validating genome-wide CRISPR-Cas9 function improves screening in the oleaginous yeast *Yarrowia lipolytica*, *Metabolic Engineering* (2019), doi: https://doi.org/10.1016/j.ymben.2019.06.007.

Validating genome-wide CRISPR-Cas9 function improves screening in the oleaginous yeast *Yarrowia lipolytica*

Cory Schwartz[1], Jan-Fang Cheng[2], Robert Evans[2], Christopher A. Schwartz[3], James M. Wagner[4], Scott Anglin[5], Adam Beitz[5], Weihua Pan[6], Stefano Lonardi[6], Mark Blenner[5], Hal S. Alper[4,7], Yasuo Yoshikuni[2], and Ian Wheeldon[1]*

1. Chemical and Environmental Engineering, University of California Riverside, Riverside, CA 92521
2. DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Walnut Creek, CA 94598
3. Department of Civil and Mechanical Engineering, United States Military Academy, West Point, NY 10996
4. McKetta Department of Chemical Engineering, The University of Texas at Austin, 200 E Dean Keeton St. Stop C0400, Austin, TX 78712
5. Chemical and Biomolecular Engineering, Clemson University, Clemson, SC 29634
6. Computer Science and Engineering, University of California Riverside, Riverside, CA 92521 USA
7. Institute for Cellular and Molecular Biology, The University of Texas at Austin, 2500 Speedway Avenue, Austin, TX 78712

*Correspondence: iwheeldon@engr.ucr.edu

Abstract

Genome-wide mutational screens are central to understanding the genetic underpinnings of evolved and engineered phenotypes. The widespread adoption of CRISPR-Cas9 genome editing has enabled such screens in many organisms, but identifying functional sgRNAs still remains a challenge. Here, we developed a methodology to quantify the cutting efficiency of each sgRNA in a genome-scale library, and in doing so improve screens in the biotechnologically important yeast *Yarrowia lipolytica*. Screening in the presence and absence of native DNA repair enabled high-throughput quantification of sgRNA function leading to the identification of high efficiency sgRNAs that cover 94% of genes. Library validation enhanced the classification of essential genes by identifying inactive guides that create false negatives and mask the effects of successful disruptions. Quantification of guide effectiveness also creates a dataset from which determinants of CRISPR-Cas9 can be identified. Finally, application of the library identified novel mutations for metabolic engineering of high lipid accumulation.

Introduction

A critical challenge in CRISPR-based library screens is the inability to separate active from inactive guides to ensure genome-wide coverage and generate high confidence hits. Typically, multiple guide RNAs are designed to target each gene of interest with the hypothesis that some guides may not be functional (Joung et al., 2017; Li et al., 2014). This strategy maximizes the likelihood of full genome coverage through redundant targeting of each gene and has been successful in identifying new phenotypes (Koike-Yusa et al., 2014; Sidik et al., 2016), but the presence of inactive guides can obscure screening results and create false negatives (Ong et al., 2017). While *in silico* design and activity predictions are emerging, complete training sets that correlate CRISPR endonuclease activity, guide sequence, and local genetic context are not yet available.

Herein, we demonstrate a methodology to measure and validate the activity of each guide RNA in a genome-wide library for screening in the oleaginous yeast *Yarrowia lipolytica.* We selected this non-conventional yeast because it has value as a bioprocessing host for the conversion of biomass derived sugars and industrial waste streams (*e.g.*, glycerol, alkanes, and fatty acids) into value added chemicals and fuels (Qiao et al., 2017; Schwartz et al., 2017a; Wagner et al., 2018). Unlike the model yeast *Saccharomyces cerevisiae*, DNA repair in *Y. lipolytica* and most other eukaryotes is dominated by nonhomologous end-joining (NHEJ) (Lobs et al., 2017). We reasoned that the effectiveness of each guide RNA in a high coverage CRISPR library could be quantified by comparing library evolution in an NHEJ-proficient strain and an NHEJ-deficient strain. In a strain lacking NHEJ and a homologous template (*i.e.,* NHEJ-disrupted haploid *Y. lipolytica*), double strand break repair is severely limited, and the most likely outcome of a Cas9-induced DNA double strand break is cell death (Schwartz et al., 2017b). Using this approach, CRISPR-Cas9 activity can be coupled to cell viability thus providing a facile, quantitative metric of guide efficiency.

Validated CRISPR libraries promise to generate more accurate and robust genetic screens with a drastically reduced false negative rate compared with libraries that are naïve to the cutting efficiency of each guide. We demonstrate this by identifying essential genes and by screening for the industrially relevant phenotype of increased lipid accumulation. In addition, when coupled with genome structure analysis (*i.e.*, nucleosome occupancy), validation experiments can provide insights into the biological effects that dictate guide RNA effectiveness and genome accessibility. Defining a list of validated guide RNAs for any given organisms is valuable metabolic engineering resource in and of itself, and is also an important step in moving towards high coverage, combinatorial screens that simultaneously target multiple genes in the genome (Shen et al., 2017).

Results

Genome-wide CRISPR-Cas9 library design, construction, and analysis

We designed a library of single guide RNAs (sgRNAs) to target 7,854 coding sequences in *Y. lipolytica* PO1f (Figure 1A, Fig. S1A, and Table S1). Unique sgRNAs were designed to target the first 300 exon base pairs in each gene, scored, and then ranked based on their predicted on-target cutting efficiency (Doench et al., 2014). The first 300 bp were targeted to maximize the likelihood of disrupting gene function in the case of an indel. The complete library was subsequently designed to contain the 6 highest scoring sgRNAs for each gene, along with 480 nontargeting controls. The library was synthesized and subsequently cloned into an expression vector with sgRNA expression driven by a synthetic RNA polymerase III (Pol III) promoter (Schwartz et al., 2017c; Schwartz et al., 2016). Sequencing revealed that over 97% of the designed sgRNAs were well represented in the library (Fig. S1B).

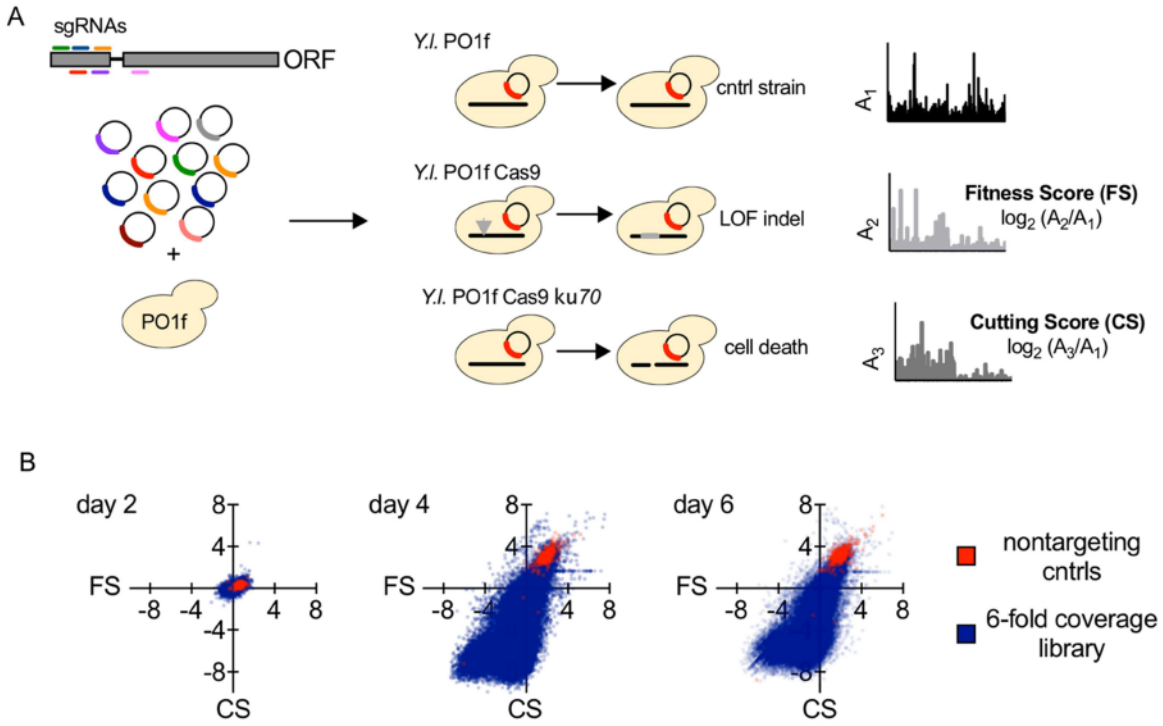


Figure 1. Genome-wide CRISPR-Cas9 validation and screening in *Yarrowia lipolytica*. (A) Schematic representation of sgRNA library design and workflow for genome-wide validation and screening. *Y. lipolytica* PO1f was used as the base strain for all experiments. Fitness score (FS) experiments used PO1f with constitutive expression of Cas9 from *Streptococcus pyogenes* (PO1f Cas9). Cutting score (CS) experiments used PO1f Cas9 with functionally disrupted nonhomologous end-joining (NHEJ) by inactivation of *KU70* (PO1f Cas9 *ku70*). LOF indel indicates loss of function insertion or deletion, and cntrl strain indicates the negative control strain that does not express Cas9 but does contain the sgRNA library. (B) Comparison of FS to CS for each sgRNA after 2, 4, and 6 days of growth on defined minimal media containing 2% glucose. Transformations were done in biological triplicates with at least 100-fold coverage of the library in each replicate. FS and CS values shown are the mean of the three biological replicates.

Growth screens in three different *Y. lipolytica* strains enabled us to experimentally determine two separate metrics for each sgRNA (Figure 1A). A fitness score (FS) was defined as the $\log_2$ ratio of the abundance of any given sgRNA in a Cas9 expressing strain (PO1f Cas9) to the abundance of the same sgRNA in the negative control (PO1f). Each sgRNA was also given a cutting score (CS) by calculating the same $\log_2$ ratio, but in this case comparing a Cas9 expressing NHEJ deficient strain (PO1f Cas9 *ku70*) to the control strain. FS is a measure of how harmful a CRISPR-Cas9 mediated indel was to cell growth, whereas CS is a measure of how efficient an sgRNA was for inducing a Cas9-mediated DNA double strand break. By using transformed PO1f as the denominator in FS and CS calculations, any bias introduced from variability in plasmid maintenance (due to a specific sgRNA sequence or otherwise) is implicitly accounted for, and changes in sgRNA abundance can be attributed to Cas9 function and the resulting effect on cell fitness. A pilot scale experiment with a nontargeting sgRNA and known functional sgRNAs targeting one essential gene and one nonessential gene confirmed that the outcomes of the FS and CS experiments were consistent with this scoring methodology (Fig. S2). A second pilot scale experiment analyzing a subset of sgRNAs that span the spectrum of CS

showed, as expected, that cell growth is strongly correlated with CS in the PO1f Cas9 *ku70* strain (Fig. S2). The absence of growth in cells containing an sgRNA with a low CS also suggests that alternative DNA repair pathways, such as microhomology-mediated end-joining, do not significantly contribute to repair.

The evolution of the full genome-wide library over three subculture cycles (6 days in total) and corresponding CS and FS values are shown in Figure 1B. Only after two growth cycles (day 4) were significant changes in cell phenotype and library distribution observed. The nontargeting control population behaved as expected, shifting to the upper right quadrant of the FS/CS plot indicating nonfunctional sgRNAs that resulted in no change in cell fitness. The day 4 and day 6 data indicate that the majority of the library was able to create double strand breaks. This is apparent from the shift toward the lower two quadrants. The lower left quadrant is indicative of sgRNAs that both cut (negative CS) and have a negative effect on growth (negative FS). The lower right quadrant also contains sgRNAs that cut effectively but have a neutral or positive effect on cell growth (positive FS). The upper left quadrant is only sparsely populated around the origin, likely because this quadrant represents instances with observed negative growth effects in the absence of a functional sgRNA (positive CS). Overall, three important trends are apparent: 1) the library skews strongly towards negative CS values, indicating that many of the sgRNAs are functional; 2) sgRNAs that do not produce a cut in the genome can be identified in the upper right quadrant; and, 3) a broad range of FS values results from outgrowth-based screens, suggesting that the library can be used to select for a wide variety of phenotypes. Based on these results and the observation that changes in fitness were not observed after one growth cycle (2 days), we selected day 4 data for all subsequent analysis.

sgRNA validation and analysis of functional determinants

Based on the average CS of the nontargeting controls, we set a value of 2 as the threshold to identify inactive sgRNAs (*i.e.* noncutters, CS> 2). At the other end of the spectrum, we set a CS value of -5 (a 32-fold reduction in sgRNA abundance) as the lower limit of sgRNAs considered to be excellent cutters (CS< -5). This threshold was set based on a subpopulation of sgRNAs whose abundance was reduced to zero after two subculture cycles (4 days), which corresponded to a CS value of -5 or lower. Two intermediate categories were also defined: moderate (-5<CS< -1) and poor (-1<CS< 2) cutters. Moderate and poor were separated at a CS of -1 because this value indicates at least some depletion of the sgRNA after two subculture cycles. Figure 2A shows the CS distribution of the full library as well as six subpopulations ranked from best sgRNA per gene (1st) to worst sgRNA per gene (6th). Each subpopulation represents 1-fold coverage of the genome. The best subpopulation contained an excellent cutter for 94.6% of the targeted genes, while 82.6% of the genes have at least two excellent cutters. Histograms of each subpopulation, and the number of excellent cutters per gene are presented (Figures 2B and C, and Fig. S3).

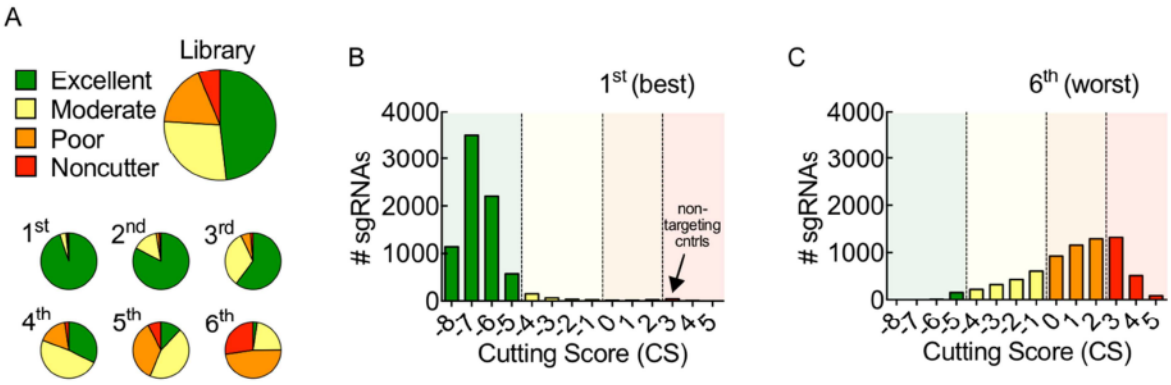


Figure 2. Genome coverage and distribution of CRISPR-Cas9 cutting efficiency. (A) Library and genome coverage by sgRNA classification: Excellent (CS<-5; green; 48% of full library), Moderate (-5<CS< -1; yellow; 28%), Poor (-1<CS< 2; orange; 18%), and Noncutter (>2; red; 6%). The sgRNAs for each gene were ranked by CS from 1st to 6th and separated into six subpopulations each representing 1-fold genome coverage. (B,C) Histograms of the 1st and 6th sgRNA subpopulations. Mean CS values were determined after two growth cycles (4 days) from triplicate biological samples.

The analysis of CS across the library revealed several trends in sequence determinants of active sgRNAs. For example, a total of 1,808 sgRNAs were found to contain a polyT motif (four or more consecutive "T" bases). This subpopulation had a significantly higher average CS than the library as a whole (CS_{polyT} = -1.389, n = 1,808 and $CS_{library}$ = -3.786, n = 46,234; t-test $p < 0.0001$, t = 30.43, df = 48,040). The apparent inefficiency of sgRNAs containing a polyT motif is not unexpected, as polyT sequences are known terminators for RNA Pol III. As such, it is likely that sgRNAs containing a polyT motif have poor functional expression causing reduced CRISPR-Cas9 activity. The effect of RNA secondary structure was also examined, but was found to be minimal (Fig. S5).

The position of sgRNAs within each chromosome was also found to have an effect. sgRNAs that target close to chromosomal ends were found to have increased cutting scores in comparison to the full library (Figure 3A and Fig. S4). Previous studies have reported that genes near the ends of chromosomes can be transcriptionally silenced by the chromatin structure of telomeres, thus providing the likely mechanism of reduced CRISPR-Cas9 activity at chromosomal ends observed here (Doheny et al., 2008; Gottschling et al., 1990).

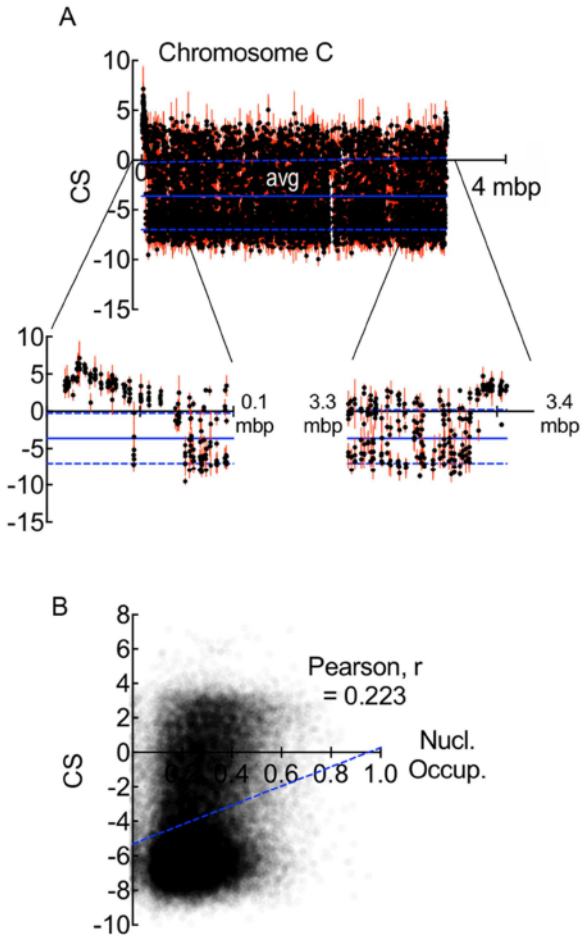


Figure 3. Effects of chromosome position and chromatin structure on CRISPR-Cas9 activity. (A) Cutting score of all sgRNAs targeting chromosome C, plotted by position. The terminal 100 kbp of both ends are shown expanded. Each data point represents the mean and standard deviation of biological triplicates of CS calculated at day 4. The solid blue line represents the average CS for the full sgRNA library. The standard deviation of the CS library average is shown as blue dashed lines. Chromosomes A, B, D, E, and F are presented in Supplementary Fig. 4. (B) Relationship between nucleosome occupancy and cutting score. Each data point represents the mean CS of biological triplicates at day 4 and the relative average nucleosome occupancy for each (n=45,247). Pearson's correlation is 0.223.

Given this result, we also elected to study the effect of genome-wide chromatin structure on sgRNA activity. The CS values of all sgRNAs in the library were compared to experimental measurements of nucleosome occupancy across the genome (Tsankov et al., 2010). As shown in Figure 3B, nucleosome occupancy positively correlated with CS. These data reveal that occupancy provides a stronger correlation to CRISPR-Cas9 activity (Pearson's, r = 0.223) than the algorithm used to design the library (Doench et al., 2014) (r = 0.0525) or other algorithms (Doench et al., 2016; Guo et al., 2018; Labuhn et al., 2018) (r = 0.0511; r = 0.1431, r=0.0365; Fig. S6).

Activity validated CRISPR-Cas9 library improves essential gene analysis

Inactive sgRNAs in CRISPR libraries can produce false negatives in essential gene screening (Ong et al., 2017). This problem arises when one or more poor cutting sgRNAs mask the effect of successful disruptions on the FS average for a particular gene. The data on sgRNA activity provided by the CS experiments can eliminate this issue by focusing analysis on the validated cutters and excluding inactive sgRNAs. Figure 4A shows the rank-ordered FS calculated from the full library, which is naïve to the effectiveness of each sgRNA and includes all excellent, moderate, poor, and noncutter sgRNAs for each gene (the naïve library). Twelve nonessential genes and twelve genes that are known to be essential across eukaryotes are shown as two subpopulations (Figure 4 and Table S2). The two subpopulations partially overlapped and had closely related FS distributions when calculated from the naïve library (Figures 4A and B). Three of the genes are notable; *ACT1*, *MYO1*, and *FOL2* are genes critical to eukaryotic cell viability, but were not distinguishable from nonessential genes, suggesting that the presence of poor cutting sgRNAs artificially increased their FS. Analysis with only excellent cutters (the validated library) resulted in a clear separation between the subpopulations: *ACT1*, *MYO1*, and *FOL2* clustered as expected with other essential genes, and the difference between FS of the essential and nonessential subpopulations increased significantly (Figures 4C and D).

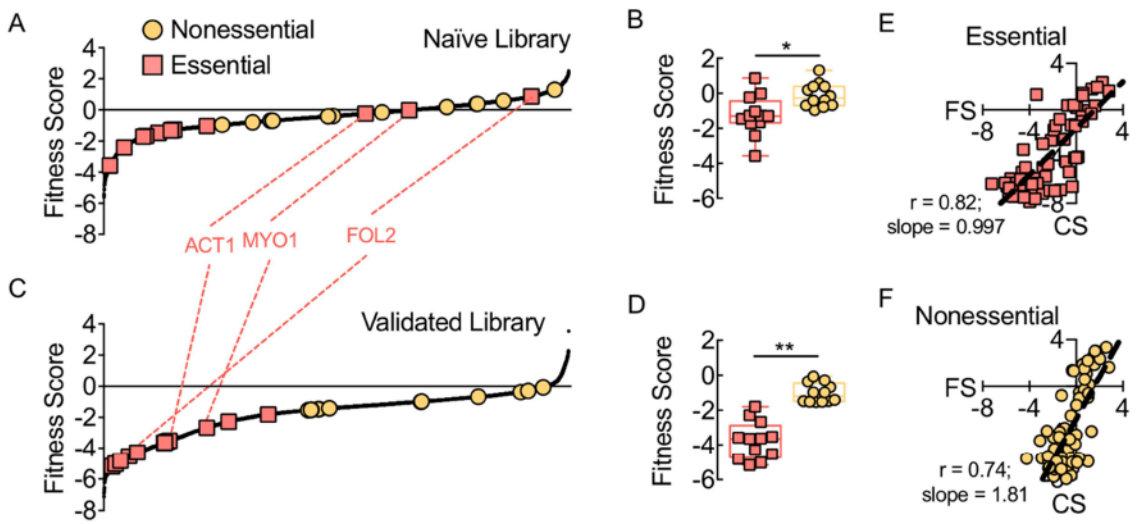


Figure 4. Library validation enhances essential gene identification. (A) Rank-ordered fitness score (FS) for each gene calculated using the full 6-fold sgRNA library (naïve library). Twelve nonessential (yellow) and putatively essential (red) genes are shown. (B) FS for essential and nonessential genes using the naïve library. Boxes extend from the 25th to 75th percentiles, the box line indicates the median, and the whiskers extend from the 10th to the 90th percentiles, n = 12 for both nonessential and essential. Comparison between the means was accomplished by two-sided t-test (t = 3.029 and p = 0.0062). Statistical significance is shown graphically with * indicating p<0.01 and ** indicating p<0.0001. (C) Rank-ordered FS for each gene calculated using only the validated excellent sgRNAs (validated library). For the 5.4% of genes that did not have an excellent sgRNA, the best cutter was included. (D) FS for essential and nonessential gene subpopulations using the validated library (t = 7.764 and p < 0.0001). (E,F) Comparison of FS to CS for each of the sgRNAs targeting the 12 selected essential and nonessential genes. Pearson's coefficients from linear regression (r) and slopes are shown (n = 72). FS and CS values are the

mean of three biological replicates, the standard deviations associated with the means are presented in Fig. S7.

One method of evaluating sgRNA effectiveness is to compare multiple targets to the same known essential gene (Doench et al., 2016; Xu et al., 2015). By definition, disruptions to essential genes are fatal (with or without intact DNA repair), which leads to a hypothesis similar to that of the CS experiments. That is, double strand breaks lead to cell death, therefore viability in a growth screen can be used as a measure of cutting efficiency. This hypothesis is supported by the observed trends in the essential and nonessential sgRNA subpopulations (Figures 4E and F). The regression line of CS/FS data for all 6 of the sgRNAs targeting the curated essential gene subpopulation has a slope of 0.997 ($r = 0.82$), suggesting a quantitative correlation between FS and CS, and providing experimental evidence to support CS as a metric for CRISPR-Cas9 activity. Appropriately, this relationship does not hold for the nonessential gene subpopulation.

The validated library, containing only highly active sgRNAs, identified 1,377 genes as essential in *Y. lipolytica*, which represents 17.5% of coding sequences (Figure 5 and Table S3; see Methods for essential gene identification methodology). Essential genes were identified by comparison to known essential and nonessential gene subpopulations. Using the same statistical tests, the naïve library (containing both active and inactive sgRNAs) identified only 4.5% of genes (359 genes) as essential. Genome-wide knockout collections and extensive experimental evidence in model yeasts suggest that ~20% of yeast genes are essential (19.5% or 1,268 genes in *S. cerevisiae* (Cherry, 2015) and 26.1% or 1,260 genes in *Schizosaccharomyces pombe* (Kim et al., 2010)), indicating that the naïve library produced a high number of false negatives in our *Y. lipolytica* essential gene screening. A recent transposon-based analysis in *Y. lipolytica* supports our results (Patterson et al., 2018). The transposon screen identified 22.4% of *Y. lipolytica* genes as essential, an increase in 586 genes over the CRISPR-based analysis in this work. The validated CRISPR library implicitly accounts for genomic targets that may be inaccessible due to heterochromatin structure, a known experimental limitation of transposon-based screens that can cause false positives (Gangadharan et al., 2010). The difference between these data sets is also likely due to challenges in transposon targeting of short genes (Wang et al., 2018) (Fig. S8).

To further confirm essential gene classification, we repeated the analysis with different known essential gene subpopulations as training sets. Specifically, we compared essential gene lists using the validated library (as described above) with a similar analysis using 186 ribosomal genes as the training set. Using the ribosomal genes as a basis for comparison resulted in a *Y. lipolytica* essential gene list of 1,549 genes, an increase of 172 genes over those identified using the curated 12-gene training set shown in Figure 4. Combing the two training sets resulted in near identical analysis to the ribosomal-only training set; 1,548 genes were identified as essential (Table S3). Because of limitations in ribosomal gene identification and uncertainty in the essentially of genes classified as ribosomal in *Y. lipolytica*, here we report the 1,377 genes defined as essential based on the curated essential subpopulation. The curated set of 12 genes includes genes with a range of cellular functions that are supported by previously published experiments (Table S2).

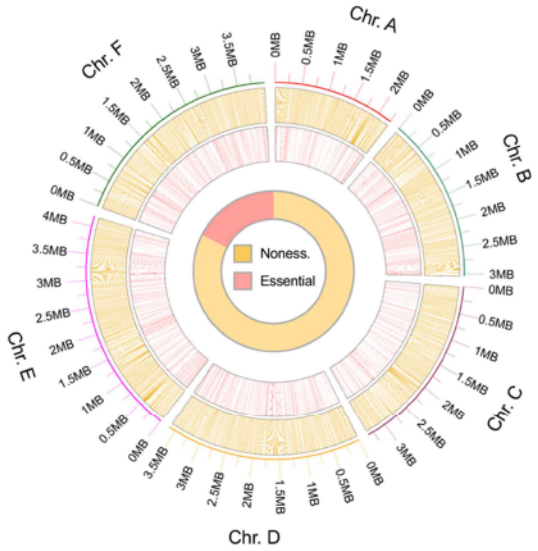


Figure 5. *Y. lipolytica* genome map. *Y. lipolytica* genome map with all nonessential and essential genes when grown on synthetic minimal media with 2% glucose at 30 °C are shown by chromosomal position. Nonessential genes (n = 6,477; yellow) are shown in the outer ring, and essential genes (n = 1,377; red) are shown in inner ring.

Of the 1,377 genes identified as essential for growth of *Y. lipolytica*, 961 had a homolog in *S. cerevisiae*, 480 of which were found to be essential in both organisms (Fig. S9). One hundred fifty-one genes essential in *Y. lipolytica* are conditionally essential in *S. cerevisiae* for respirative growth. Because *Y. lipolytica* is an obligate aerobe, respiration is essential and therefore the gene conditionally essential in *S. cerevisiae* were expected to be essential in *Y. lipolytica*. Ninety-five more genes are duplicated in the *S. cerevisiae* genome, and 69 are involved in amino acid or nucleotide biosynthesis. Gene ontology (GO) term analysis (Ashburner et al., 2000) was also performed on all genes that had a *S. cerevisiae* homolog. Genes involved in transcription, translation, cellular organization, and cell cycle were found to have significantly lower FS (Fig. S9).

To complete the essential gene analysis in *Y. lipolytica*, we compared the 416 genes that were found to be essential but had no *S. cerevisiae* homolog against the genomes of the fission yeast *Schizosaccharomyces pombe* and the filamentous fungi *Aspergillus nidulans* (Table S4). One hundred and ninety-eight of the 416 genes in this subpopulation have a homolog in *S. pombe*, 111 of which are essential to *S. pombe* viability. The comparison to *A. nidulans* was not as comprehensive because only limited information on the essential genes of this species have been reported. However, 248 genes (of the 416 gene without a homolog in *S. cerevisiae*) were identified in *A. nidulans*, none of which are known to be essential in *A. nidulans*.

Genome-wide screening for growth and nongrowth associated phenotypes

To demonstrate the utility of the CRISPR library, we performed a series of growth and nongrowth associated screens to identify genes linked to new phenotypes. The validation experiments described above revealed that over 94% of genes in the genome were targeted by highly functional sgRNAs (*i.e.*, "Excellent" sgRNAs), thus providing confidence in genome-wide screening and limiting the likelihood of false negatives. As an initial test, we conducted the classic genetic screen for canavanine resistance. Canavanine is a molecule that is structurally similar to arginine and is toxic to yeast unless the *CAN1* gene is disrupted. Ninety-four percent of

sequenced colonies that survived the canavanine challenge (50 mg/L over 2 days of growth) harbored an expression vector encoding a *CAN1* targeting sgRNA (Figure 6A). The remaining sgRNA targeted YALI1_E36540g, a vacuolar phosphate transporter. All identified sgRNAs were excellent cutters.

In a second screen, *Y. lipolytica* strains with increased lipid staining were isolated from the CRISPR library using a fluorescent lipid dye and fluorescence activated cell sorting (Figure 6B). sgRNA sequences targeting four different genes were associated with a substantial increase in lipid staining (defined as 3 standard deviations or more above the mean fluorescence intensity for the unsorted strain). All hits were produced with sgRNAs classified as "Excellent". Two of the four identified genes (YALI1_A22440g and YALI1_F16045g) have known roles in lipid turnover and growth on fatty acids in other yeast species (Henry et al., 2012; Strijbis et al., 2009). A third gene (YALI1_F31153g) is likely involved in the enzymatic processing of acyl-CoA species (Sherman et al., 2004), a function that is often intertwined with lipid metabolism (*e.g.*, β-oxidation). The final gene (YALI1_F20943g) in the set has no known function or similarity to a well-annotated sequence, but two different excellent sgRNAs targeting the gene were isolated from the high lipid staining subpopulation. This redundant sgRNA targeting supports the validity of the target and demonstrates the utility of the screen for identifying novel genotype to phenotype relationships.

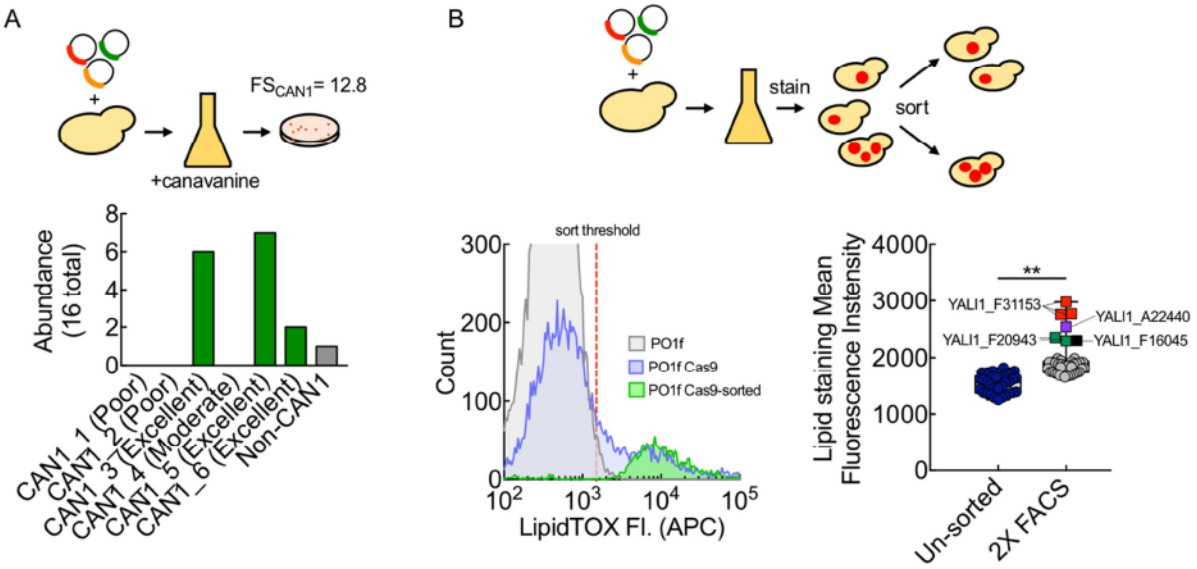


Figure 6. Genome-wide CRISPR-Cas9 screens for growth and non-growth associated phenotypes. (A) Growth screen for canavanine resistance. Cells were grown in the presence of 50 mg L^{-1} canavanine for 3 days, plated and sequenced. (B) Two populations were stained and analyzed using fluorescence activated cell sorting (PO1f with library; red in histogram, and PO1f Cas9 with library; blue in histogram). The PO1f Cas9 strain with library was sorted twice for high lipid producers to yield a high lipid producing population (2x FACS Pool; yellow in histogram). Several strains from the high lipid producing population were isolated and compared to a mock/un-sorted control consisting of cells subjected to the same process but without a fluorescence sorting cutoff. Comparison between the means was accomplished by two-sided t-test ($n = 34, 41$; $t = 6.713$; and, $p < 0.0001$). Statistical significance is shown graphically with * indicating $p<0.01$ and ** indicating $p<0.0001$.

Discussion

With genome-scale mutagenesis technologies such as MAGE, TRMR, Tn-seq and other similar approaches, it is now possible to generate and track genome edits and their effects on native and engineered phenotypes (Si et al., 2017; van Opijnen et al., 2009; Wang et al., 2009; Warner et al., 2010). Genome-wide CRISPR systems are complementary to these approaches (Hart et al., 2015; Koike-Yusa et al., 2014; Sidik et al., 2016; Wang et al., 2015) and can have the added functionality of genome targeted site saturation mutagenesis (Bao et al., 2018; Garst et al., 2017; Guo, 2018). However, these techniques have not yet been widely translated to non-model and other industrially and medically relevant organisms. The widespread adoption of CRISPR–Cas9 promises to enable genome-wide engineering and screening for a vast array of hosts (Löbs et al., 2017; Lobs et al., 2017; Nodvig et al., 2015; Shan et al., 2013; Sidik et al., 2016). Here, we designed and applied such a system for functional genomic screening and non-native phenotype engineering in the yeast *Y. lipolytica*. Key to the success of the system was our ability to functionally disable native nonhomologous DNA repair, which allowed quantification of the cutting efficiency of each sgRNA-Cas9 complex in parallel and yielded a list of validated sgRNAs for nearly every gene in the *Y. lipolytica* genome.

The genome-wide data created by the cutting score (CS) experiments also generates new information that can be used to identify the determinants of CRISPR-Cas9 function, as well as improve screen accuracy. Our experiments in *Y. lipolytica* revealed that 48% of guides in the 6-fold coverage library were highly active, and that 94.6% of protein coding sequences were targeted by at least one excellent cutting sgRNA (Figure 2). Similar data has recently been generated using genome-wide CRISPR-Cas9 systems in *E. coli* (Guo et al., 2018; Gutierrez et al., 2018). This bacterium natively lacks DNA repair by NHEJ creating experiments similar to those conducted here with disrupted *KU70* function. Overall, library performance in *E. coli* was similar to that observed in *Y. lipolytica*; ~44% of *E. coli* sgRNAs had at least 32-fold depletion[22], compared with ~48% in *Y. lipolytica* (Figure 2). However, an sgRNA activity prediction algorithm derived from the *E. coli* data had poor correlation with the *Y. lipolytica* CS data (r=0.0365; Fig. S9D) demonstrating that CRISPR-Cas9 activity can vary significantly across organisms and expression systems(Guo et al., 2018).

Library validation also substantially improved essential/nonessential gene classification (Figures 4 and 5) by significantly reducing false negatives that arise from analysis with all active and inactive guides in the library. A nongrowth associated screen was also successful in identifying novel gene disruptions for high lipid accumulation (Figure 6). Correlation of CS data with nucleosome occupancy and several sgRNA scoring algorithms revealed that occupancy was a stronger predictor of function than *in silico* design (Figure 3; see Refs. (Doench et al., 2016; Doench et al., 2014; Horlbeck et al., 2016; Labuhn et al., 2018; Tsankov et al., 2010)). We also observed that chromosomal ends were difficult to target. Current scoring algorithms have been limited by training set quality and size, relying on subsets of essential genes and/or high expression reporters to learn about the relationships between guide RNA sequence and resultant Cas9 activity (Doench et al., 2016; Doench et al., 2014; Xu et al., 2015). The genome-wide activity measurements obtained in the CS experiments provide a substantially richer data set for correlating guide sequence parameters to system-level characteristics. We envision that our experiments in *Y. lipolytica* may also be possible in other eukaryotes by genetically disrupting or chemically inhibiting native nonhomologous DNA repair. As such, the approach described here not only resulted in a new validated screening tool for *Y. lipolytica*, but also represents a generalizable methodology for enhancing CRISPR-based screens in other organisms.

Materials and Methods

Strains and media

All strains were derived from *Yarrowia lipolytica* strain PO1f (MatA, leu2-270, ura3-302,xpr2-322, axp-2). The PO1f Cas9 strain was constructed by markerless integration of a UAS1B8-TEF(136)-Cas9-CycT expression cassette into the A08 site (Blazeck et al., 2011; Schwartz et al., 2017c). The PO1f Cas9 *ku70* strain was constructed by disrupting *KU70* using CRISPR-Cas9 as previously described (Schwartz et al., 2016).

Yeast culturing was done at 30 °C, unless specifically noted otherwise. Yeast strains were grown in YPD (1% Bacto yeast extract, 2% Bacto peptone, 2% glucose) for nonselective growth. Cells transformed with sgRNA-expressing plasmids were selected in media deficient in leucine (SD-leu; 0.67% Difco yeast nitrogen base without amino acids, 0.069% CSM-leu (Sunrise Science, San Diego, CA), and 2% glucose). For canavanine resistance screening, 50 µg/mL was added to SD-leu. For lipid overproduction experiments, SDL-leu (0.67% Difco yeast nitrogen base without amino acids, 0.069% CSM-leu, and 8% glucose) was used for outgrowth.

Plasmid construction

For integration of Cas9, pIW1009 (UAS1B8-Cas9 integration into A08 site) was constructed. pHR_A08_hrGFP (Addgene #84615) was digested with BssHII and NheI, and Cas9 was inserted via Gibson Assembly after PCR via Cr_1250 and Cr_1254 from pCRISPRyl (Addgene #70007). To facilitate cloning of the library of sgRNAs, pIW771 (SCR1'-tRNA-AvrII site) was generated by digesting pCRISPRyl with BamHI and HindIII and circularizing.

sgRNA library design

Custom Matlab scripts were used to design each sgRNA in the library based on a recent assembly of the *Y. lipolytica* PO1f parent strain genome, W29 (Magnan et al., 2016). sgRNAs were designed for all non-redundant protein coding sequences. The top 6 sgRNAs, as ranked by on target efficiency score(Doench et al., 2014) that met the following criteria were selected for each gene: (1) uniqueness of target sequence and PAM (protospacer adjacent motif; final 12 bp + NGG) in genome, (2) located within first 300 bp of gene, and (3) target sequence present in both genome and mRNA sequence (to avoid targeting introns). Four hundred eighty sgRNAs of random sequence, confirmed to not target in the genome using the same methodology as was used for checking for uniqueness of designed sgRNAs, were included as nontargeting controls. All designed sgRNAs and corresponding data are available in Supplementary Table 1.

sgRNA library cloning

Four oligo pools (Supplementary Table 4) consisting of 25% of the designed sgRNAs each were ordered from Twist Bioscience. Equal volumes of the oligo pool (1 µM) and a complementary primer (5 µM) were mixed together and annealed using the Thermo annealing advanced protocol and the calculated Tm (melting temperature) of the complementary primers using IDT's OligoAnalyzer 3.1. For libraries 1 and 3, second strand synthesis reactions were completed using T4 DNA polymerase (NEB) and sgRNA-Rev2, gel extracted, and purified using Zymo Research Zymo-Spin 1 columns. Libraries 2 and 4 were amplified via Q5 DNA polymerase (NEB) using 60mer_pool-F and spacer-AarI.rev or pLeu-mock-sgRNA.fwd and sgRNA-Rev2 with 0.2 picomoles of DNA as template for 7 cycles and column purified.

For libraries 1, 3, and 4, Gibson Assembly HiFi HC 1-step Master Mix (SGI-DNA) were used to clone into pIW771 digested with AarI. Library 2 was digested with AarI and cloned into pIW771 digested with AarI with a GoldenGate reaction using T4 DNA ligase (NEB). Cloned DNA was transformed into NEB 10-beta *E. coli* and plated. Sufficient electroporations were

performed for each library to yield a >10-fold excess colonies for the number of library variants, and the plasmid library was isolated. Errors in sequence were found to have arisen primarily due to oligo synthesis (approximately 1 in 500 bp, as reported by the vendor) and were not dependent on cloning method.

Yeast transformation and screening

Transformation of *Y. lipolytica* with the sgRNA plasmid library was done using a previously described method, with slight modifications (Schwartz et al., 2016). Two mL of YPD was inoculated with a single colony of the strain of interest and grown in a 14 mL tube with shaking at 200 RPM for 24-30 h. Then, 1.5 mL of cells were washed with transformation buffer (0.1 M LiAc, 10 mM Tris (pH=8.0), 1 mM EDTA) via centrifugation at 4,000 *g* and resuspended in 600 µL transformation buffer. To these cells, 18 µL of ssDNA mix (8 mg/mL Salmon Sperm DNA, 10 mM Tris (pH=8.0), 1 mM EDTA), 90 µL of β-mercaptoethanol mix (5% β-mercaptoethanol, 95% triacetin), and 4 µg of plasmid library were added, mixed via pipetting, and incubated. After incubation at room temperature for 30 min, 900 µL of PEG mix (70% w/v PEG (3,350 MW)) was added and mixed via pipetting, and incubated at room temperature for 30 min. Cells were then subjected to heat shock for 25 min at 37 °C, washed with 10 mL H_2O, and used to inoculate 25 mL of SD-leu media for outgrowth experiments. Dilutions of the transformation were plated on solid SD-leu media to calculate transformation efficiency. Transformations were pooled as necessary to ensure adequate diversity to maintain library representation and to minimize the effect of plasmid instability (100x coverage, 5 x 10^6 total transformants per biological replicate).

Outgrowth was allowed to proceed in 25 mL of liquid media in a 250 mL baffled flask. After 2 days, cells reached confluency (optical density at 600 nm (OD_{600}) > 10), and approximately 150 µL (at least 200-fold library coverage) were used to inoculate 25 mL of fresh media as desired. Also at each timepoint, 1 mL of culture was taken, DNase I (New England Biolabs) and the corresponding buffer were added, and the mix was incubated for 1 h at 30 °C to digest any plasmid DNA in the media. Pellets were then frozen and stored at -80 °C for future analysis. Each growth cycle from inoculation to sampling represented approximately 7 cell doublings.

Library isolation and sequencing

Pellets were thawed and resuspended in 1 mL H_2O. Cells were split into 5 samples of 200 µL, and plasmid was isolated using a Zymo Yeast Miniprep Kit (Zymo Research). Samples from a single pellet were pooled, and plasmid copy number was quantified using real time quantitative PCR (RT-qPCR) with M13_F and M13_R and SsoAdvanced Universal SYBR Green Supermix (Biorad) and verified to be higher than 1 x 10^7.

To prepare samples for next generation sequencing, isolated plasmid was subjected to PCR using forward (Cr_1665-1668) and reverse primers (Cr_1669-1673 and Cr_1709-1711) containing all necessary sequences and barcodes (Joung et al., 2017) (Supplementary Table 4). At least 1 x 10^7 plasmids were used as template, and PCR reactions were not allowed to go to completion to avoid biased amplification. PCR product at 250 bp was gel extracted, samples were pooled at equimolar amounts, and submitted for sequencing on a NextSeq500 at the UCR IIGB core facility.

Next generation sequencing results were demultiplexed and mapped to each sgRNA using custom scripts. A total of 453 sgRNAs were not present in the sequencing data. In addition, 440 sgRNAs had only a very low number of reads (less than 10) in the untransformed library or were erroneously designed, and were excluded from downstream analysis. Pairwise

comparison between normalized read abundances for biological replicates were done to verify consistency (Fig. S10).

Cutting and fitness score analysis

To calculate fitness score (FS) and cutting score (CS) for each replicate, the number of reads for each sgRNA in each sequencing sample was normalized to the total number of reads, with a pseudo-count of 0.5 added when no reads were mapped to a given sgRNA for a given biological replicate. Normalized read counts for biological triplicates were then averaged together. The $\log_2$ for each abundance was then taken, and then the FS was calculated by subtracting the PO1f $\log_2$ abundance from the PO1f Cas9 $\log_2$ abundance. CS was calculated in analogous way, but with PO1f Cas9 replaced with PO1f Cas9 *ku70*.

sgRNA analysis

Analysis of sgRNA characteristics was done using a range of publicly available tools and data. Nucleosome occupancy was determined by mapping sgRNAs to previously published nucleosome occupancy data (Tsankov et al., 2010), adding the occupancy for each base pair in the target sequence, and dividing by 20. The scoring for predicting on-target activity for each sgRNA was carried out using previously published algorithms (Doench et al., 2016; Doench et al., 2014; Labuhn et al., 2018). Secondary structure was predicted using the default settings for RNA using Quikfold on the DINAMelt web server (http://unafold.rna.albany.edu/) (Zuker, 2003). polyT sequences were identified by searching each sgRNA sequence for 4 consecutive "T". sgRNAs at chromosome ends were annotated as the 50 sgRNAs nearest either end of each chromosome, except for the end of chromosome A (84 sgRNAs), the beginning of chromosome C (126 sgRNAs), and the end of chromosome E (71 sgRNAs). Ends of chromosomes determinations were informed by where sgRNA cutting scores increased.

Essential gene identification

Essential genes were identified based on FS at day 4. A two-sided t-test was used to determine which genes had a FS different than the 12 putative nonessential genes ($p<0.05$). A second two-sided t-test was also used to determine genes with a FS lower than a numerical cutoff defined by the average FS of the 12 essential gene training set plus two standard deviations (1.02 for naïve library, -1.57 for validated library; see Tables S2 and S3). Multiple comparisons were accounted for using a Bonferroni correction as necessary. Genes found to be significant by both tests were classified as essential genes. The essential gene analysis was repeated for the validated library using ribosomal genes in place of the 12 essential gene training set (cutoff for t-test = -0.87), and combining ribosomal genes with the 12-gene training set (cutoff for t-test = -0.91).

Essential gene comparison to *S. cerevisiae*

Mapping of all *Y. lipolytica* genes to *S. cerevisiae* genes was carried out using BLAST ($E<10^{-30}$). Essential genes in *S. cerevisiae* were those annotated as being "inviable" null mutants (Cherry et al., 2012).

Canavanine screen

Glycerol stocks of the transformed library taken from day 2 cultures were grown in SD-leu for 2 days, and then used to inoculate SD-leu with 50 µg/mL canavanine. Cultures were then allowed to grow to confluency and plated on solid SD-leu. Single colonies were subjected to colony PCR using Cr_1742 and Cr_1743 to identify the sgRNA expressed in each colony.

LipidTOX Deep Red Staining, Fluorescence Activated Cell Sorting (FACS), and Flow Cytometry Confirmation

CRISPR libraries were cultivated for FACS in media and conditions that were previously established for *Y. lipolytica* lipid characterization(Wagner et al., 2018). Specifically, cultures were started from frozen glycerol stocks (PO1f with CRISPR library and PO1f Cas9 with CRISPR library, starting from day 2 stocks) by inoculating 50 mL of SD-leu media in a 250 mL baffled shake flask, followed by 2 days of outgrowth at 28°C. The resulting inoculum cultures were then used to start 50 mL lipid production cultures in SDL-leu media (initial optical density of 0.1) with 80 g/L glucose in 250 mL shake flasks. These shake flask cultures were grown using air-permeable plugs by shaking for 3 days at 28°C and 225 RPM in a shaking incubator (New Brunswick Scientific I 26). HCS LipidTOX Deep Red (Thermo Fisher Scientific) was then used to fluorescently stain lipids according to the manufacturer-recommended protocol: cells were pelleted, all culture media was removed, and then the cell pellet was re-suspended in 100 µL of a 1:200 dilution of lipid stain in phosphate buffered saline (PBS). Two successive rounds of fluorescence activated cell sorting (FACS) were performed on a BD FACS Aria III using the APC laser/filter set. For high lipid sorting, the top 0.3% of $3x10^6$ APC^+ cells and the top 3% of $1x10^5$ APC^+ cells were collected. After the first APC sort, the collected cells were outgrown in SD-leu to saturation (3 days) and glycerol stocked. Fresh cultures were then started from the frozen glycerol stock by inoculating 50 mL of SD-leu media in a 250 mL baffled shake flask, followed by 2 days of outgrowth at 28°C. The resulting inoculum culture was then used to start a 50 mL lipid production culture in SDL-leu media with 80 g/L glucose (initial optical density of 0.1) in a 250 mL shake flask. This shake flask culture was grown using air-permeable plugs by shaking for 3 days at 28°C and 225 RPM in a shaking incubator (New Brunswick Scientific I 26) until ready for the second sort. After the second APC sort, the collected cells were outgrown in SD-leu to saturation (3 days) and glycerol stocked again. A mock-sorted / un-sorted pool was also generated in parallel throughout this double sorting process by passing the same pool of cells through the same FACS process/conditions without applying the sorting gate (*i.e.*, all cells were collected, regardless of APC fluorescence).

From the 2X sorted (2X FACS) glycerol stock and mock/un-sorted glycerol stock, individual colonies were then isolated on SD-leu plates to isolate individual strains. 48 colonies from the 2X FACS and 48 colonies from the mock/un-sorted plates were then subjected to colony PCR and Sanger sequencing of the sgRNA cassette. Clones with sgRNAs that were successfully colony PCR amplified, sequenced, and identified as cutters (see cutting and fitness score analysis above) were subsequently re-screened to establish a lipid staining mean fluorescence intensity for individual clones (41 from 2X FACS, 34 from mock/unsorted) using the same procedure as for FACS, but with a BD LSRII Fortessa analytical flow cytometer for quantifying APC channel fluorescence.

Statistical Analysis

Data is shown as mean ± standard deviation, with n values indicated in the figure legends. P values were generated using the statistical approach described in the figure legends using either Graphpad Prism 7 or R 3.3.2. To determine the correlation between different sgRNA metrics (such as nucleosome occupancy or design algorithm score), Graphpad Prism 7 was used to calculate the Pearson's correlation. Comparisons between FS from essential and nonessential genes determined from the two different sgRNA libraries was done using a two-sided t-test, implemented in Graphpad Prism 7. Essential genes were identified by using a two-sided t-test to compare the FS of each gene to a nonessential gene reference group (to identify those which were significantly different). A second two-sided t-test was used to compare the FS

of each gene to a numerical cutoff, which was determined by the FS of a known essential gene reference group. Multiple comparisons were accounted for using Bonferroni correction.

Acknowledgments

This work was supported by DOE Joint Genome Institute grant CSP-503076 (IW, HA, and MB) NSF 1706545 (IW). HA and JW were supported by the Office of Naval Research (ONR) under grant N00014-15-1-2785. JW acknowledges additional support from the National Science Foundation (NSF) Graduate Research Fellowship Program (DGE-1110007). SA, AB, and MB were supported by NSF 1706134. SL and WP were supported by NSF 1526742. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231.

Author contributions

CS and IW conceived the study. CS, CAS, and IW designed the library. JFC, RE, and YY constructed the library. CS, WP, SL, and IW performed library validation. CS, SA, AB, MB, and IW designed, completed and analyzed the essential gene experiments. CS, JW, HA, and IW completed and analyzed phenotypic screens. All authors wrote and edited the manuscript.

Data and materials availability

All Illumina sequencing reads have been deposited in the Sequence Read Archive under project accession PRJNA478042.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25, 25-9.

Bao, Z., HamediRad, M., Xue, P., Xiao, H., Tasan, I., Chao, R., Liang, J., Zhao, H., 2018. Genome-scale engineering of Saccharomyces cerevisiae with single-nucleotide precision. Nature biotechnology. 36, 505-508.

Blazeck, J., Liu, L. Q., Redden, H., Alper, H., 2011. Tuning Gene Expression in Yarrowia lipolytica by a Hybrid Promoter Approach. Appl. Environ. Microbiol. 77, 7905-7914.

Cherry, J. M., 2015. The Saccharomyces Genome Database: Advanced Searching Methods and Data Mining. Cold Spring Harbor protocols. 2015, pdb prot088906.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Simison, M., Weng, S., Wong, E. D., 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic acids research. 40, D700-D705.

Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., Virgin, H. W., Listgarten, J., Root, D. E., 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature biotechnology. 34, 184-+.

Doench, J. G., Hartenian, E., Graham, D. B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B. L., Xavier, R. J., Root, D. E., 2014. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nature biotechnology. 32, 1262-U130.

Doheny, J. G., Mottus, R., Grigliatti, T. A., 2008. Telomeric Position Effect-A Third Silencing Mechanism in Eukaryotes. Plos One. 3.

Gangadharan, S., Mularoni, L., Fain-Thornton, J., Wheelan, S. J., Craig, N. L., 2010. DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo. Proceedings of the National Academy of Sciences of the United States of America. 107, 21966-21972.

Garst, A. D., Bassalo, M. C., Pines, G., Lynch, S. A., Halweg-Edwards, A. L., Liu, R. M., Liang, L. Y., Wang, Z. W., Zeitoun, R., Alexander, W. G., Gill, R. T., 2017. Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. Nature biotechnology. 35, 48-55.

Gottschling, D. E., Aparicio, O. M., Billington, B. L., Zakian, V. A., 1990. Position Effect at Saccharomyces-Cerevisiae Telomeres - Reversible Repression of Pol-Ii Transcription. Cell. 63, 751-762.

Guo, J. H., Wang, T. M., Guan, C. G., Liu, B., Luo, C., Xie, Z., Zhang, C., Xing, X. H., 2018. Improved sgRNA design in bacteria via genome-wide activity profiling. Nucleic acids research. 46, 7052-7069.

Guo, X., Chavez, A., Tung, A., Chan, Y., Kaas, C., Yin, Y., Cecchi, R., Garnier, S.L., Kelsic, E.D., Schubert, M., DiCarlo, J.E., Collins, J.J., Church, G.M., 2018. High-throughput creation and functional profiling of DNA sequence variant libraries using CRISPR-Cas9 in yeast. Nature biotechnology. 36, 540-546.

Gutierrez, B., Wong Ng, J., Cui, L., Becavin, C., Bikard, D., 2018. Genome-wide CRISPR-Cas9 screen in *E. coli* identifies design rules for efficient targeting. biorXiv.

Hart, T., Chandrashekhar, M., Aregger, M., Steinhart, Z., Brown, K. R., MacLeod, G., Mis, M., Zimmermann, M., Fradet-Turcotte, A., Sun, S., Mero, P., Dirks, P., Sidhu, S., Roth, F. P., Rissland, O. S., Durocher, D., Angers, S., Moffat, J., 2015. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell. 163.

Henry, S. A., Kohlwein, S. D., Carman, G. M., 2012. Metabolism and regulation of glycerolipids in the yeast Saccharomyces cerevisiae. Genetics. 190, 317-49.

Horlbeck, M. A., Witkowsky, L. B., Guglielmi, B., Replogle, J. M., Gilbert, L. A., Villalta, J. E., Torigoe, S. E., Tjian, R., Weissman, J. S., 2016. Nucleosomes impede Cas9 access to DNA in vivo and in vitro. Elife. 5.

Joung, J., Konermann, S., Gootenberg, J. S., Abudayyeh, O. O., Platt, R. J., Brigham, M. D., Sanjana, N. E., Zhang, F., 2017. Genome-scale CRISPR-Cas9 knockout and transcriptional activation screening. Nature Protocols. 12, 828-863.

Kim, D. U., Hayles, J., Kim, D., Wood, V., Park, H. O., Won, M., Yoo, H. S., Duhig, T., Nam, M., Palmer, G., Han, S., Jeffery, L., Baek, S. T., Lee, H., Shim, Y. S., Lee, M., Kim, L., Heo, K. S., Noh, E. J., Lee, A. R., Jang, Y. J., Chung, K. S., Choi, S. J., Park, J. Y., Park, Y., Kim, H. M., Park, S. K., Park, H. J., Kang, E. J., Kim, H. B., Kang, H. S., Park, H. M., Kim, K., Song, K., Song, K. B., Nurse, P., Hoe, K. L., 2010. Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. Nature biotechnology. 28, 617-623.

Koike-Yusa, H., Li, Y. L., Tan, E. P., Velasco-Herrera, M. D., Yusa, K., 2014. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nature biotechnology. 32, 267-273.

Labuhn, M., Adams, F. F., Ng, M., Knoess, S., Schambach, A., Charpentier, E. M., Schwarzer, A., Mateo, J. L., Klusmann, J. H., Heckl, D., 2018. Refined sgRNA efficacy prediction improves large- and small-scale CRISPR-Cas9 applications. Nucleic Acids Research. 46, 1375-1385.

Li, W., Xu, H., Xiao, T. F., Cong, L., Love, M. I., Zhang, F., Irizarry, R. A., Liu, J. S., Brown, M., Liu, X. S., 2014. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol. 15.

Löbs, A. K., Engel, R., Schwartz, C., Flores, A., Wheeldon, I., 2017. CRISPR-Cas9-enabled genetic disruptions for understanding ethanol and ethyl acetate biosynthesis in Kluyveromyces marxianus. Biotechnology for Biofuels. 10.

Lobs, A. K., Schwartz, C., Wheeldon, I., 2017. Genome and metabolic engineering in non-conventional yeasts: Current advances and applications. Synthetic and Systems Biotechnology. 1-10.

Magnan, C., Yu, J., Chang, I., Jahn, E., Kanomata, Y., Wu, J., Zeller, M., Oakes, M., Baldi, P., Sandmeyer, S., 2016. Sequence Assembly of Yarrowia lipolytica Strain W29/CLIB89 Shows Transposable Element Diversity. Plos One. 11, e0162363.

Nodvig, C. S., Nielsen, J. B., Kogle, M. E., Mortensen, U. H., 2015. A CRISPR-Cas9 System for Genetic Engineering of Filamentous Fungi. Plos One. 10.

Ong, S. H., Li, Y., Koike-Yusa, H., Yusa, K., 2017. Optimised metrics for CRISPR-KO screens with second-generation gRNA libraries. Sci Rep-Uk. 7.

Patterson, K., Yu, J., Landberg, J., Chang, I., Shavarebi, F., Bilanchone, V., Sandmeyer, S., 2018. Functional Genomics for the Oleaginous Yeast Yarrowia Lipolytica. Metab Eng.

Qiao, K. J., Wasylenko, T. M., Zhou, K., Xu, P., Stephanopoulos, G., 2017. Lipid production in Yarrowia lipolytica is maximized by engineering cytosolic redox metabolism. Nature Biotechnology. 35, 173-177.

Schwartz, C., Frogue, K., Misa, J., Wheeldon, I., 2017a. Host and Pathway Engineering for Enhanced Lycopene Biosynthesis in Yarrowia lipolytica. Front Microbiol. 8.

Schwartz, C., Frogue, K., Ramesh, A., Misa, J., Wheeldon, I., 2017b. CRISPRi repression of nonhomologous end-joining for enhanced genome engineering via homologous recombination in Yarrowia lipolytica. Biotechnology and Bioengineering. 114, 2896-2906.

Schwartz, C., Shabbir-Hussain, M., Frogue, K., Blenner, M., Wheeldon, I., 2017c. Standardized Markerless Gene Integration for Pathway Engineering in Yarrowia lipolytica. ACS synthetic biology. 6, 402-409.

Schwartz, C. M., Hussain, M. S., Blenner, M., Wheeldon, I., 2016. Synthetic RNA Polymerase III Promoters Facilitate High-Efficiency CRISPR-Cas9-Mediated Genome Editing in Yarrowia lipolytica. ACS synthetic biology. 5, 356-359.

Shan, Q. W., Wang, Y. P., Li, J., Zhang, Y., Chen, K. L., Liang, Z., Zhang, K., Liu, J. X., Xi, J. J., Qiu, J. L., Gao, C. X., 2013. Targeted genome modification of crop plants using a CRISPR-Cas system. Nature biotechnology. 31, 686-688.

Shen, J. P., Zhao, D. X., Sasik, R., Luebeck, J., Birmingham, A., Bojorquez-Gomez, A., Licon, K., Klepper, K., Pekin, D., Beckett, A. N., Sanchez, K. S., Thomas, A., Kuo, C. C., Du, D., Roguev, A., Lewis, N. E., Chang, A. N., Kreisberg, J. F., Krogan, N., Qi, L., Ideker, T., Mali, P., 2017. Combinatorial CRISPR-Cas9 screens for de novo mapping of genetic interactions. Nat Methods. 14, 573-+.

Sherman, D., Durrens, P., Beyne, E., Nikolski, M., Souciet, J. L., 2004. Genolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts. Nucleic acids research. 32, D315-8.

Si, T., Chao, R., Min, Y. H., Wu, Y. Y., Ren, W., Zhao, H. M., 2017. Automated multiplex genome-scale engineering in yeast. Nature communications. 8.

Sidik, S. M., Huet, D., Ganesan, S. M., Huynh, M. H., Wang, T., Nasamu, A. S., Thiru, P., Saeij, J. P. J., Carruthers, V. B., Niles, J. C., Lourido, S., 2016. A Genome-wide CRISPR Screen in Toxoplasma Identifies Essential Apicomplexan Genes. Cell. 166, 1423-+.

Strijbis, K., van Roermund, C. W., Hardy, G. P., van den Burg, J., Bloem, K., de Haan, J., van Vlies, N., Wanders, R. J., Vaz, F. M., Distel, B., 2009. Identification and characterization of a complete carnitine biosynthesis pathway in Candida albicans. Faseb J. 23, 2349-59.

Tsankov, A. M., Thompson, D. A., Socha, A., Regev, A., Rando, O. J., 2010. The Role of Nucleosome Positioning in the Evolution of Gene Regulation. Plos Biol. 8.

van Opijnen, T., Bodi, K. L., Camilli, A., 2009. Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat. Methods. 6, 767-U21.

Wagner, J. M., Williams, E. V., Alper, H. S., 2018. Developing a piggyBac Transposon System and Compatible Selection Markers for Insertional Mutagenesis and Genome Engineering in Yarrowia lipolytica. Biotechnol J. 13, e1800022.

Wang, H. H., Isaacs, F. J., Carr, P. A., Sun, Z. Z., Xu, G., Forest, C. R., Church, G. M., 2009. Programming cells by multiplex genome engineering and accelerated evolution. Nature. 460, 894-U133.

Wang, T., Birsoy, K., Hughes, N. W., Krupczak, K. M., Post, Y., Wei, J. J., Lander, E. S., Sabatini, D. M., 2015. Identification and characterization of essential genes in the human genome. Science. 350, 1096-1101.

Wang, T. M., Guan, C. G., Guo, J. H., Liu, B., Wu, Y. A., Xie, Z., Zhang, C., Xing, X. H., 2018. Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. Nature communications. 9.

Warner, J. R., Reeder, P. J., Karimpour-Fard, A., Woodruff, L. B. A., Gill, R. T., 2010. Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. Nature biotechnology. 28, 856-U138.

Xu, H., Xiao, T. F., Chen, C. H., Li, W., Meyer, C. A., Wu, Q., Wu, D., Cong, L., Zhang, F., Liu, J. S., Brown, M., Liu, X. S., 2015. Sequence determinants of improved CRISPR sgRNA design. Genome Res. 25, 1147-1157.

Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic acids research. 31, 3406-3415.

Highlights

- A library of highly functional Cas9 sgRNA for >94% of *Y.lipolytica* genes was identified
- Disruption of native DNA repair enabled quantitative analysis of sgRNA function
- Chromatin structure and chromosomal position significantly influenced sgRNA activity
- Validation of sgRNA function improved essential gene classification
- Functional genomic screening revealed 4 gene knockouts for lipid overproduction