UNIVERSITY OF CALIFORNIA
RIVERSIDE

Development of CRISPR Based Synthetic Biology Tools for Genome Engineering and Functional Genomic Screening in the Industrially Relevant Oleaginous Yeast *Yarrowia lipoytica*

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Chemical and Environmental Engineering

by

Adithya Ramesh

December 2022

Dissertation Committee:
Dr. Ian Wheeldon, Chairperson
Dr. Yanran Li
Dr. Robert Jinkerson

The Dissertation of Adithya Ramesh is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the help and support from many people. First and foremost, I would like to express my heartfelt gratitude to my advisor, Prof. Ian Wheeldon for his mentorship, support and confidence in me throughout the course of my doctoral career. Ian, I will always be grateful to you for seeing the potential in me, encouraging me to do my PhD and working to secure the funding to make it a reality. Our exhaustive research discussions and your support of my research ideas have made me the independent researcher I am today, and I will always hold you in the highest regard for your support and encouragement during difficult times.

I am also extremely grateful to all my lab mates, current and alum, who have made the last six years of my life a time to cherish. Thank you all for your help, guidance, countless hours of research discussions, and most importantly for providing a warm and welcoming atmosphere. Cory and Ann-Kathrin, thank you both especially for mentoring me in the lab, and always making time to answer incessant queries from a young and curious researcher.

The text of this dissertation is in part, a reprint of the material as it appears in the journals of ACS Synthetic Biology (2020, chapter 2), Methods in Molecular Biology (2021, chapter 3), and Nature Communications (2022, chapter 4). The corresponding author Ian Wheeldon, listed in those publications, directed and supervised the research, which forms the basis for this dissertation. None of this work would have been possible

without our wonderful collaborators. I would like to express my most sincere appreciation for Dr. Stefano Lonardi, our collaborator and the Principal Investigator of the Lonardi lab in the Computer Science and Engineering Department, who co-supervised and co-wrote the research article that appears in chapter 4. I would also like to thank Dipankar Baisya who developed the computational framework for the work listed in chapter 4 and co-wrote the manuscript. My heartfelt gratitude to Varun Trivedi who developed the analysis framework for the research in chapter 5 and co-wrote the manuscript which is under consideration for publication in Communications Biology. Thank you to Dr. Cory Schwartz who helped perform experiments that contributed to the success of the work listed in chapters 4 and 5. Thank you also to Thomas Ong, Jessica Adams, Jamie Garcia, Aida Tafrishi, and Amirsadra Mohseni for their contributions, either experimental or computational that contributed to the research that appears in chapters 2, 4 and 5. I am indebted to Peter Sheffield and Jeff Braman from Agilent technologies for their help and contributions in the form of printed DNA libraries, that were instrumental in conducting the research that appears in chapters 4 and 5. My sincere gratitude to Dr. Yanran Li and Dr. Robert Jinkerson for serving on my thesis committee, for their encouragement, and for their insightful questions and comments during my advancement to doctoral candidacy, which set me on path I am today. I would also like to acknowledge UC Riverside, and the National Science Foundation for supporting the work in this thesis.

I could not have undertaken this journey without the love and support of my mother and grandparents. Mum, words cannot express how thankful I am for the tremendous support, love and hope you've given me throughout my life. I would not be the man I am

today without you. My warm and most heartfelt gratitude to my grandparents who were nearly as invested in my PhD as I was. Without the strength I gained from my family, my doctoral career would not have been possible.

Last but certainly not the least, I would like to express my deepest gratitude to my closest friends, Aditya, Anudeep, Pranjal, Aashreth, Badri, Shraavan, and Siddarth who buoyed me up, kept me sane and motivated me throughout the course of my PhD. Thank you all for your unwavering support and confidence in me, and for letting me unwind and enjoy myself so I could continue my work energized. I am blessed to have you all in my life as my friends.

ABSTRACT OF THE DISSERTATION

Development of CRISPR Based Synthetic Biology Tools for Genome Engineering and Functional Genomic Screening in the Industrially Relevant Oleaginous Yeast *Yarrowia lipoytica*

by

Adithya Ramesh

Doctor of Philosophy, Graduate Program in Chemical and Environmental Engineering
University of California, Riverside, December 2022
Dr. Ian Wheeldon, Chairperson

Microbial biochemical production as a renewable alternative to traditional methods is a rapidly growing sector of industrial biotechnology. Non-conventional microbes are attractive targets for metabolic engineering to produce biochemicals as they can present a range of desirable traits that may help avoid complex and intensive engineering of less suitable model hosts. *Yarrowia lipolytica* is one such non-conventional yeast with an abundant acetyl-CoA pool and native capacity to produce and accumulate lipids to high levels. While there have been significant advances in the metabolic engineering of this yeast for the biosynthesis of oleochemicals and other value-added products, there is also a dearth of synthetic biology tools for genome engineering, functional genomic screening and rapid strain development. We have sought to overcome these limitations by developing CRISPR-Cas9 and Cas12a systems for multiplexed gene knockout, integration, regulation, and genome-wide screening. However, prediction of highly active guide RNA (gRNA) which are crucial in effective genome editing and improving confidence in hit calling, remains a challenge. To address this, we constructed two genome-wide libraries, one using

SpCas9 and the other using LbCas12a, to target all protein coding sequences. A negative selection screen in the absence of DNA repair, was used to generate gRNA activity scores for both endonucleases. This genome-wide data served as input to a deep learning algorithm, DeepGuide, that could accurately predict high activity gRNA for both Cas9 and Cas12a. Another critical challenge in accurately assessing screening outcomes is accounting for the variability in gRNA activity. Poorly active guides targeting genes essential to screening conditions obscure the growth defects that are expected from disrupting them. Thus, we also developed acCRISPR, an end-to-end pipeline that used gRNA activity scores to provide an activity correction to the screening outcomes, thus accurately determining the fitness effect of disrupted genes. acCRISPR analysis of the Cas9 and Cas12a screens in Yarrowia enabled the determination of a high-confidence set of essential genes for growth under glucose, a common carbon source used for the industrial production of oleochemicals. acCRISPR was also used in high salt and low pH tolerance screens, to identify known and novel genes related to stress tolerance. Collectively, this thesis presents an experimental-computational framework for CRISPR-based functional genomics studies that may be expanded to other non-conventional organisms of interest.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1: Introduction

1.1 Background

1.1.1 Industrial Biotechnology: Scope and challenges

Since the early 2000s, the acceptance of climate change and the limited availability of fossil fuels have driven researchers towards the pursuit of energy independence, greenhouse gas mitigation, and sustainable and renewable alternatives to chemical production[1–3]. Extensive research and development efforts have been carried out to advance technologies for biomass production, conversion and valorization to produce biofuels and other value added bioproducts[4]. As a result of this rapid growth, the biotechnology sector now contributes over $388 billion towards annual revenues in the US, representing more than 2% of the US the gross domestic product (GDP)[5]. A wide variety of industrial biochemicals such as biofuels, biopolymers, nutraceuticals, food additives, oleochemicals, flavors and fragrances, and other specialty chemicals have contributed to nearly $150 billion of these revenues. Biochemical production has thus become the fastest growing subsector with consistent year over year growth averaging nearly 10% in the last decade[5].

Microbial biosynthesis of biofuels, commodity and high-value specialty biochemicals has garnered a lot of attraction in the past decade for many reasons[6]. (i) Microbes may be engineered to grow on waste products from several industrial processes such as crude glycerol, molasses or lignocellulosic biomass; (ii) bioproduction can be achieved with short process cycles due to short doubling times of microbes; (iii) microbial production, which

can be scaled up in industrial bioreactors, is typically independent of climate or seasonality unlike plant or animal based sources of biochemicals; (iv) land area for microbial production is also lesser than that required for plant or animal growth, and does not compete with food production; (v) the remarkable advancement of synthetic biology and metabolic engineering has provided the technical capabilities to rapidly design and create microbial cell factories with optimized pathways for the production of a desired bioproduct[6,7]; (vi) the advent of -omics based platforms (genomics, transcriptomics, proteomics), has made large scale and high throughput characterization of microbial metabolic pathways became possible, allowing researchers identify novel targets to enhance the production of target chemicals[8–12]. Thus, a microbe may be engineered to utilize inexpensive feedstocks as carbon source to produce a value-added product of interest, even one non-native to its metabolism, to be scaled up to meet market demand as necessary.

However, the ability to design low-cost bioprocesses has not kept pace with the technological advancements in this field (CRISPR based genome editing[13], low cost of DNA synthesis and sequencing enabling high throughput -omics analyses[14]). Current outstanding challenges limiting the competitiveness of bioprocessing include expensive feedstocks or poor productivity on inexpensive feedstocks, high energy and water use, loss of productivity due to contamination, and high downstream separation costs[15]. These challenges have proven arduous to tackle in part due to the unsuitability of commonly used model organisms that play host to these bioprocesses.

A valuable approach to metabolic engineering is identifying organisms with desirable phenotypes and developing new synthetic biology tools to enhance these phenotypes. Historically, this has proven true as bioprocesses for chemical production have relied on a microorganism's ability to overproduce a specific product of interest. For example, the yeast *Saccharomyces cerevisiae* finds widespread use for the production of bioethanol, with over 30 million gallons produced worldwide in 2019[16]. This yeast remains the microbe of choice to produce ethanol, due to its ability to favor the fermentation sugars under aerobic conditions over the production of biomass, as well as its tolerance to high tires of ethanol (over 120 g/L)[3,17,18]. Other notable examples of organisms that gained industrial relevance due to native bioproduction capacity include the fungus *Penicillium chrysogenum* for the production of the antimicrobial penicillin, and filamentous fungus *Aspergillus niger* for the production of preservative citric acid[19,20].

Due to its importance in ethanol production *S. cerevisiae* has been extensively studied and developed as a model organism. The genetics, physiology, and metabolism of this host is well studied and characterized and there exist an abundance of synthetic biology tools for metabolic engineering towards specialty chemicals production. However, while this yeast is the perfect candidate for ethanol production, commercial production of alternate bioproducts (for example oleochemicals, or carotenoids) have found limited success due to limitations of the native metabolic pathways. Besides, this yeast is mesophilic, does not tolerate environmental stresses, and shows limited native or engineered capacity for growth on varied carbon sources other than glucose[21,22]. While efforts to address these limitations have been made easy by the available tools for genome engineering, success has been

limited as the phenotypes in question are often defined by complex genetic interactions that are difficult to reproduce *de novo* in a different host.

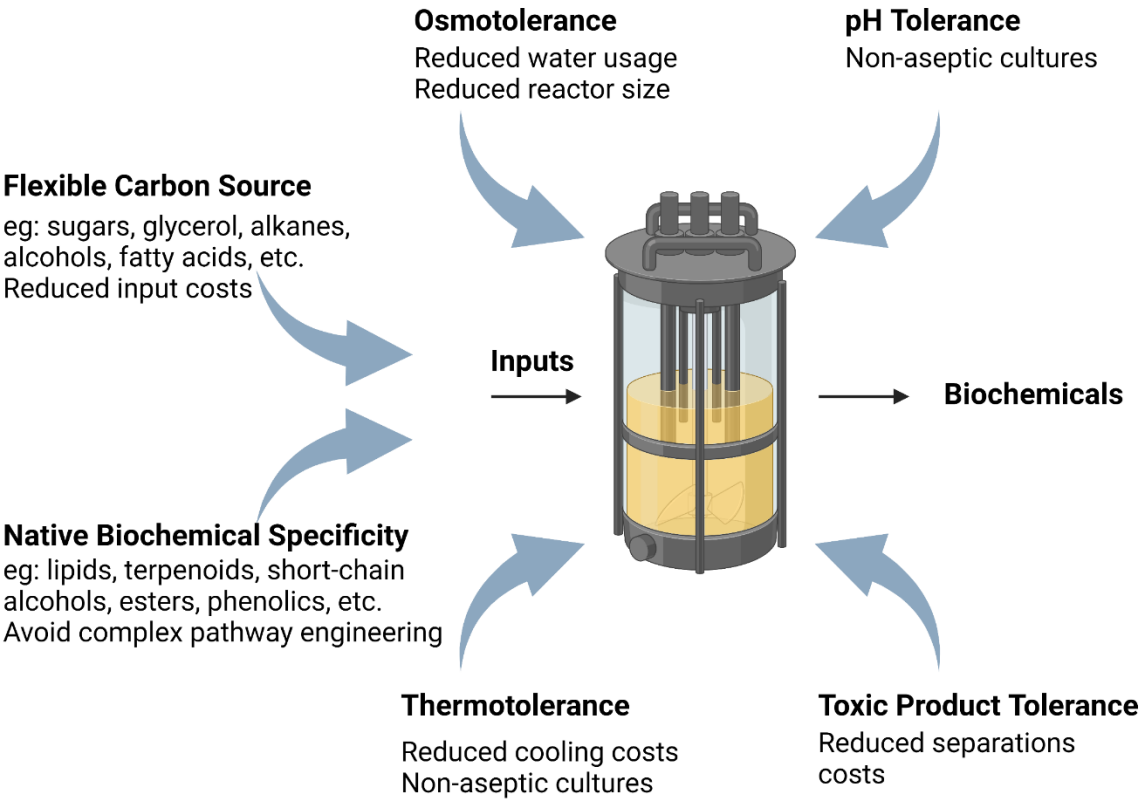

Fig 1.1. Phenotypes portrayed by non-model microorganisms that meet the needs for scalable and economically favorable industrial biochemical synthesis. In addition to native capacity of an organism to produce a desired product, other attractive features such as tolerance to various stresses and utilization of a range of substrates as carbon source, will help reduce input and process costs, and make for more economically viable bioprocesses. **(*Adapted and modified with permission from Thorwall et al. 2020*[15])**

The success of translating lab scale biochemical production to industrial scale production lies in the associated process economics. Methods and inputs that may be a nonissue at lab scale such as substrate cost, temperature control, culture sterility, water usage etc., become critical bottlenecks at scale[15]. Non model organisms natively possessing

favorable phenotypes that can address these issues at the outset, show promise of enabling the next generation of bioprocessing (Fig. 1.1). For example, extremely fast growth rates may help production hosts outcompete contaminating species, limiting losses in productivity and allowing for non-aseptic cell culturing. This is a trait exemplified by the non-conventional yeast *Kluyveromyces marxianus* touted as the fastest growing eukaryote[23]. In addition, this yeast also shows high native capacity for the production of ethyl acetate which finds widespread use as a solvent[24].

Costs associated with high water usage (as an example, ethanol plants use five times the amount of treated water to the ethanol product[25]) may be reduced by the native capacity for growth in untreated water sources, such as ocean water. The yeast *Debaromyces hansenii* has recently garnered attention as production host for its osmo-, halo- and xero-tolerance[26]. This species has also shown utility in the production of food ingredients and nutraceuticals such as xylitol, arabitol and riboflavin[26]. The oleaginous yeast *Yarrowia lipolytica* that is well known for its broad substrate utilization and its ability to accumulate lipids, also has isolates that show tolerance to salt concentrations of over 10% in solution[27]. Cooling costs also make up a significant portion of plant operation expenditure. Since microbial growth is exothermic, bioreactor cooling becomes critical to maintain viable temperatures, especially in tropical regions. Microbes such as *K. marxianus, Hansenula polymorpha*, and *Thermoanaerobacterium saccharolyticum*, are promising hosts capable of growth under temperatures as high as 50 °C while also showing ability to ferment pentose sugars which are cheaply and abundantly available from lignocellulosic biomass[28–31].

Given that the goal of biochemical production is to achieve high titers and yields, tolerance to the toxicity of either the final bioproduct or any intermediates is also an important trait. Accumulation of such products like solvents (e.g., ethanol, butanol) or organic acids (e.g., acetic acid, citric acid) or other bioactive compounds can disrupt protein folding, destabilize cellular membranes, or even disrupt DNA replication. Concentrated product streams also reduce costs associated with product separation and recovery[15]. The yeast *S. cerevisiae* is already known to tolerate high concentrations of ethanol. Another notable example is the bacterium *Clostridium acetobutylicum*, which has high tolerance to industrial solvents such as acetone, butanol and ethanol, and has been used for the production of the same[32,33]. Finally, tolerance to extreme pH conditions can allow for non-aseptic cell cultures. In addition, these organisms may be able to grow better on lignocellulosic biomass which are usually acidic due to the pretreatment process[34]. The yeast species *Issatchenkia orientalis* and *Y. lipolytica* which can grow under pH conditions less than 3.5 have thus garnered attention as potential industrial hosts. To be certain, there is unlikely to be a single host capable of addressing all bioprocessing challenges, however, expanding the number of viable hosts and building synthetic biology tools with which to take advantage of their natively favorable phenotypes, will help match microbial hosts with process needs.

1.1.2 CRISPR based synthetic biology tools for genome engineering, transcriptional control and forward genetic screening

While non model microbes do show a broad range of industrially favorable traits as discussed in the previous section, their genetics and metabolic pathways are not as well

characterized as model microbial species such as *E. coli* or *S. cerevisiae*. As well, there is typically a dearth of sophisticated synthetic biology tools that are needed for facile genome engineering in these hosts. Thus, the development of such tools is central to efforts that attempt to engineer these microbes into production strains. The advent of CRISPR technology that were discovered at the turn of the last decade has revolutionized the concept of genome engineering (Fig 1.2).

The first CRISPR endonuclease adopted for targeted double stranded break (DSB) in a wide range of organisms was the Cas9 nuclease (~1400 aa) isolated from the bacterium *Streptococcus pyogenes*. Cas9 is an RNA guided endonuclease that functions forming a ribnucleoprotein complex of a CRISPR RNA (crRNA or spacer; 20 bp in length) and a structural component (tracrRNA or transactivating crRNA; 88 bp in length) that enables complexation of the crRNA with the CRISPR associated endonuclease (i.e., Cas9). Targeting is achieved by the complementarity of the crRNA to a desired genomic locus, which must be adjacent to a protospacer adjacent motif (PAM; 'NGG' found immediately 3' of the targeted region) to activate endonuclease function. Once Cas9 is bound to the target DNA, two of its domains HNH and RuvC are responsible for a blunt end cleavage for strand complementary and non-complementary to the spacer[13,35]. For genome editing purposes, typically the tracrRNA is fused to the 3' end of the crRNA to create a short or single guide RNA (sgRNA).

A few years later, a second Cas RNA guide endonuclease slightly smaller than Cas9 (~1200 aa) was identified and termed as Cas12a[36]. While the broad function of Cas12a was

similar to that of Cas9 in effecting DSB, there were a few key differences in its mechanism of action. Cas12a recognized a T rich PAM sequence ('TTTV' found immediately 5' of the target locus), created DSB with sticky ends, and did not require a tracrRNA element. Instead, complexation of spacer to Cas12a was achieved with the help of a short 19 bp sequence called a direct repeat (DR). Most importantly, the Cas12a protein contained an endoribonuclease domain that could recognize and cleave the crRNA transcripts at 5' end of the DR sequence[37,38]. For genome editing purposes, typically the DR sequence is fused to the 5' end of the crRNA to create a sgRNA. Leveraging the Cas12a endoribonucleolytic activity at the site of the DR, multiple sgRNA may be tiled consecutively for Cas12a to process them into mature sgRNA. This has allowed for facile multiplexed genome editing with Cas12a based systems in a wide range of organisms[39–43].

Functional expression of CRISPR Cas9 or Cas12a systems for gene editing in yeast requires a codon-optimized Cas protein expression cassette with a nuclear localization tag such as SV40 fused to its C-terminus, as well as an sgRNA expression cassette. A nuclear localization tag is needed because *S. pyogenes* is a bacterium, and thus an unmodified Cas9 would localize to the cytosol in yeast. Expression of the sgRNA cassette has been achieved with both RNA Pol II and Pol III promoters[44,45]. Use of RNA Pol II promoters require additional flanking ribozymes for proper maturation of sgRNA as they do not behave like the mRNA typically transcribed by these promoters[46]. Functional expression of both components will induce a DSB at the target locus in the host cell that is repaired one of the native DNA repair pathways. Repair of the DSB by non-homologous end joining NHEJ commonly results in indels and gene inactivation, while providing a homology repair

template incentivizes repair of the break by homology directed repair (HDR) and allows for a desired sequence to be inserted at the cut site (Fig 1.1a). While the model yeast *S. cerevisiae* overwhelmingly performs HDR and requires very short homology arms (<50 bp) for precise gene knock ins[47,48], DNA repair in most non-conventional yeasts proceeds via NHEJ, and gene integrations require long homology arms (~1 kb) and often time inactivation of native NHEJ mechanism by the disruption of KU70 and KU80 genes[45,49–53].

The CRISPR toolbox in yeast has also been further expanded to include gene regulation in addition to gene disruptions and integration. The nuclease domains of Cas9 (HNH: D10A; RuvC: H840A) or Cas12a (RuvC: D832A) maybe be mutated to deactivate their nuclease activity (dCas) while still retaining their DNA targeting and binding activity. The dCas may then be targeted to promoter regions upstream of the transcriptions start site (TSS) such as the TATA box to sterically hinder transcription machinery from assembling. This technique called CRISPR interference (CRISPRi) is used for gene repression[54]. Fusion of transcriptional repressors such as Mxi1 or KRAB may increase the efficiency of repression[55]. Similarly, activation domains such as VP64 or VPR may also be fused to the Cas protein and targeted to the upstream activation elements of a promoter to achieve gene overexpression through CRISPR activation (CRISPRa)[56]. Transcriptional regulation using CRISPR systems opens up a new modality in non-model microbes, where native promoters, either constitutive or inducible are less well characterized than model species.

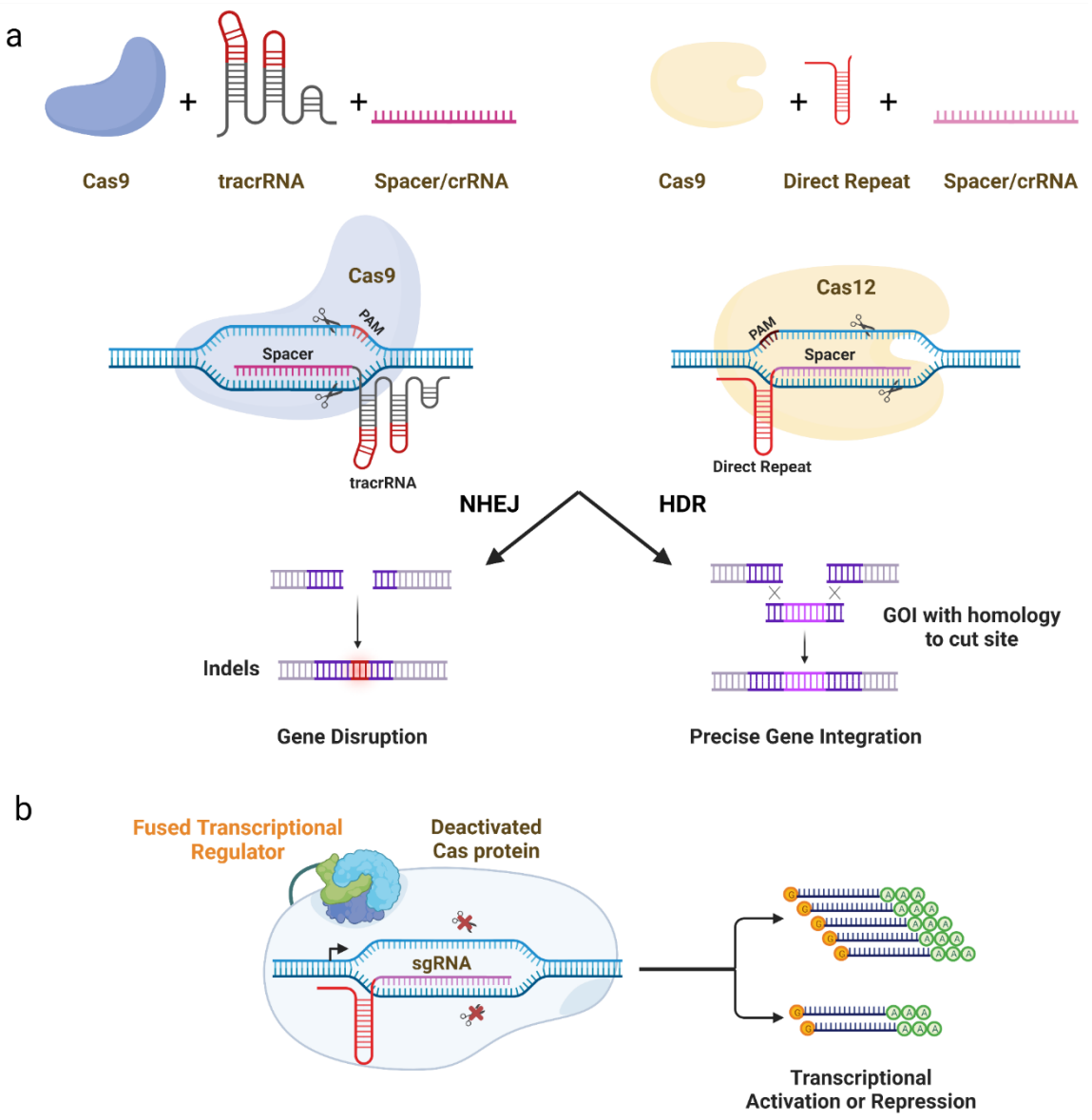

Fig 1.2. CRISPR based synthetic biology tools for gene editing and transcriptional regulation. (a) Cas9 and Cas12a based gene editing for gene disruption and gene integration. (b) Catalytically deactivated Cas proteins lacking nuclease activity may be fused with transcriptional activators (e.g., VP64, p65, Rta, among others) or transcriptional repressors (e.g., Mxi1, KRAB) to achieve gene repression or overexpression.

Another important functionality afforded by CRISPR systems is the ability to conduct forward genetic screens. While reverse genetics identifies how a specific gene is associated with a known phenotype, forward genetic analysis is an unbiased approach to uncover genes essential to defined biological phenomena. Such screens become ever more important in the context of non-model microbes, that show a wide range of interesting phenotypes, but typically lack sufficient genetic characterization of those traits. With the ability to perform targeted mutagenesis, CRISPR screens become a powerful tool for biological discovery enabling the unbiased interrogation of gene function in a genome wide scale for a wide range of applications and species[57–61].

Pooled CRISPR screens are performed by introducing various genetic perturbations (e.g., mutations in the open reading frame of every coding region in the genome) into a pool of cells (Fig. 1.3). This is generally achieved with the help of a guide RNA library such that individual cells in the pool receive different gRNAs, and subsequently edited. These mutations are then allowed to persist, and the pool of cells may also be subject to a biological challenge to favor the persistence of mutations enhancing a desired phenotype (e.g., tolerance to high temperature, high salt, low pH, toxic compounds, etc.). Subsequently, the mutants are evaluated by next generation sequencing (NGS) based counting of the gRNAs that specify each mutation. The typical results of such screens are ranked lists of genes that confer sensitivity or resistance to the biological challenge of interest. Once again, the broad utility of CRISPR systems allow the expansion of screening modalities to CRISPRi and CRISPRa for the activation and silencing of different gene targets, sometimes even in a multiplexed format[62]. One possible route forward with

CRISPR screens can take winners from the first round of screening and subject them to further rounds of screening and selection to identify mutations that when stacked enhance the desired phenotype. However, the scale and complexity of validating screening outcomes increases exponentially with each successive round of screening

While there exist other genome wide mutagenesis strategies that use Agrobacterium T-DNA, transposase mediated insertions, ethylmethyl sulfonate (EMS) mutagenesis, or UV-irradiation mutagenesis, one major advantage of CRISPR/Cas mutagenesis over such traditional methods is the causal mutations for a desired phenotype may be mapped by identifying the target gRNA sequence, followed by sanger sequencing of the target for confirmation[63–66]. Genome wide CRISPR systems are complementary to established traditional approaches and can have the added advantage of targeted site saturation mutagenesis. These techniques have not yet been widely translated to other non-model industrially relevant organisms, however the steady rise in the adoption of CRISPR systems in these organisms promises to deliver the capacity for forward genetic screening to further elucidate their genetics and metabolism.

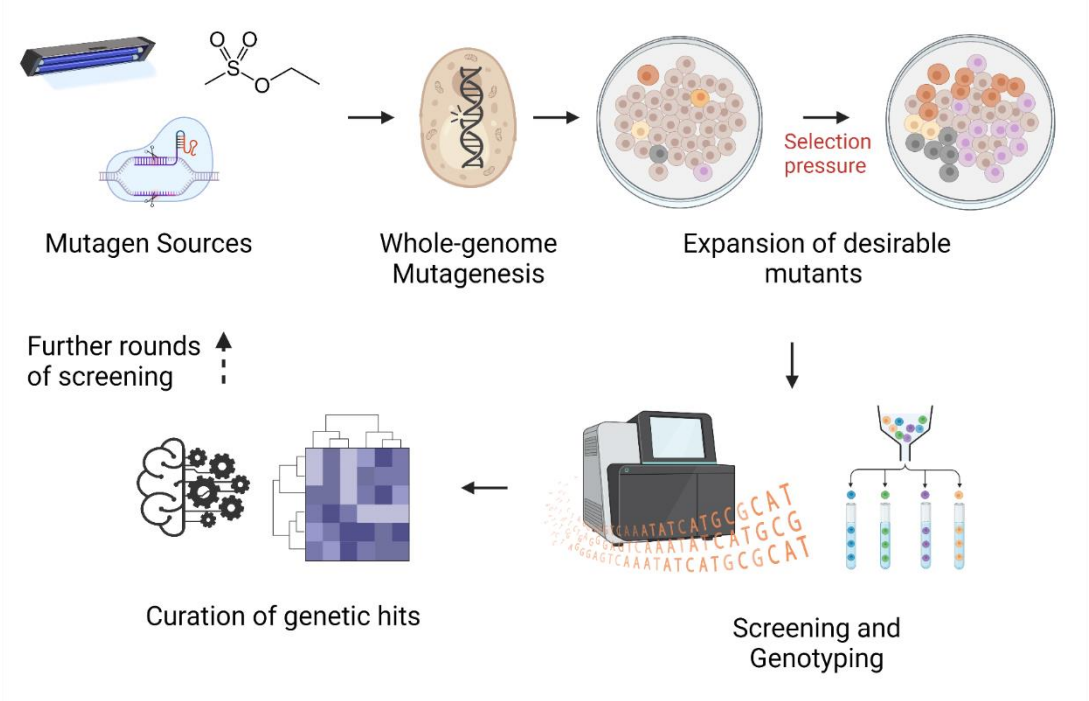

Fig 1.3. Forward genetic screening approaches to evolve desirable phenotypes and elucidate their genetic underpinnings.

1.1.3 The industrially relevant non-conventional yeast: *Yarrowia lipolytica*

Yarrowia lipolytica is a non-conventional oleaginous yeast that can utilize a wide variety of inexpensive and renewable substrates (such as sugars, glycerol, fatty acids, alkanes, and other hydrophobic substrates) as carbon sources. It also displays halotolerance and pH tolerance, with the ability to grown under high levels of salt stress (up to 10% w/v) and a wide range of pH from 4-11[27,67]. As an oleaginous yeast, *Y. lipolytica* natively accumulates up to 30% of its dry cell weight as triacylglycerides (TAGs)[68–72]. A set of gene overexpressions, deletions, and heterologous integrations have also been identified that enable this yeast to accumulate over 90% of its dry cell weight as TAGs[73]. As a consequence of this oleaginous behavior, it can accommodate a high flux of the precursor

acetyl-CoA. This yeast also has a generally regarded as safe (GRAS) status, has a fully sequenced and annotated genome and its lipid related metabolic pathways have been extensively studied[74–77].

These characteristics have made *Y. lipolytica* an attractive industrial host for the production of a wide variety of chemicals such as lipid derived biofuels and oleochemicals (fatty acid methyl and ethyl esters, fatty alkanes and alcohols, and wax esters among others), organic acids (citrate, isocitrate, α-ketoglutarate, succinate and itaconic acid), carotenoids (lycopene, β-carotene, astaxanthin), other plant terpenoids (farnesene, linalool) and sugar alcohols (such as erythritol and erythrulose)[7,78–86]. While this yeast is capable of producing a wide spectrum of biochemicals, a lot of effort has been invested in leveraging its oleaginous behavior to make it an industrial chassis for lipid biosynthesis. Traditional strategies for maximizing lipid accumulation involve, (i) deletion of TAG lipases and β-oxidation genes involved in the lipid degradation pathway, (ii) overexpression of fatty acid and TAG synthesis genes, and, (iii) minimization of flux towards competing pathways such as glycogen storage and citrate biosynthesis[87–89]. More recently, high levels of TAGs were engineered by the replacing native NAD^+ dependent enzymes with $NADP^+$ dependent variants in order to increase the cytosolic NADPH available for lipid biosynthesis. This led to a strain with a lipid productivity of 1.2 g $L^{-1}h^{-1}$, moving this process closer to industrial feasibility. Researchers from DuPont have also utilized *Y. lipolytica* as a host to produce the nutraceutical omega-3 eicosapentaenoic acid (EPA), that is sold under the commercial brand name Newharvest™ EPA oil. This was achieved by the random integration of 30

copies of nine endogenous and heterogeneous genes along with the disruption of β-oxidation, resulting in a strain accumulating EPA at 15% of its dry cell weight.

Y. lipolytica owes its success as a production host to the development of synthetic biology tools for genome engineering over the past few years. There now exist a suite of CRISPR based tools for gene editing, gene integration, and transcriptional regulation that enable facile genetic engineering[43,52,55,90–92]. However, many potential advances are needed before *Y. lipolytica* can obtain the status of a model organism. Synthetic biology tools that will facilitate multiplexed and combinatorial genome engineering strategies are necessary for shorter design-build-test-learn cycles during strain engineering. Unlike *S. cerevisiae*, *Y. lipolytica* does not have a curated list of essential genes that would ease pathway and target selection decisions for the production of novel molecules. Genome wide engineering strategies using developed the CRISPR tools provide a promising avenue for elucidating as of yet uncovered genetics and metabolism of this host. However, the field of pooled CRISPR screens is still fairly new and the experimental and bioinformatic tools to conduct and analyze such screens are still not well established in non-model hosts. The work presented in this dissertation covers the development of some of these advanced synthetic biology tools and provides biological insights and potential engineering applications using these tools. As more such tools and methods are adapted for use in non-conventional hosts, the more their relevance to industrial biotechnology will rise and the faster these microbes will become new model organisms.

1.2 Thesis organization

The work presented in this dissertation expands the development of CRISPR based synthetic biology tools for genome engineering in *Y. lipolytica*. As well, experimental and computational workflows for the implementation of forward genetic analyses using pooled CRISPR screens are established, and novel biological insights and possible applications and future directions of such methods are discussed.

Chapter 1 has introduced the concept of industrial biotechnology, its scope, potential, and outstanding challenges, and discussed how non-model organisms may play a role in tackling some of those challenges. The lack of advanced synthetic biology tools for genetic engineering in many non-model hosts and how CRISPR technologies may play a role in rectifying this issue was also briefly touched upon. Further, concepts regarding CRISPR based gene editing, transcriptional regulation, and forward genetic screening were also introduced. Finally, the utility of *Y. lipolytica* as a production host and potential advances in the toolset required to further engineering in this host was also discussed and these topics will form the basis of all other chapters in this dissertation.

In chapter 2, the expansion of the existing CRISPR synthetic biology tools in include CRIPSR-Cas12a systems is described. The ease of multiplexing for gene disruptions is shown by knocking out three genes simultaneously with high efficiency. Furthermore, gRNA length dependent cutting of the Cas12a nuclease was also showcased at a series of gRNA lengths. The lack of nuclease activity by Cas12a at gRNA lengths below 16 nt was

leveraged to introduce CRISPRi and CRISPRa modalities for gene silencing and overexpression.

Chapter 3 expands on the concept of pooled CRISPR screens further and presents detailed methodology for the design of gRNA libraries to implement genome-wide CRISPR-Cas9 and CRISPR-Cas12a screens in *Y. lipolytica*. Details regarding guide RNA uniqueness to minimize off target guide activity, as well as the design of appropriate controls for the screening experiment are also discussed. The MATLAB scripts used for the design of such CRISPR libraries are also provided here.

Genome-wide functional genetic screens have shown great success discovering genotype-phenotype relationships and in engineering new phenotypes. The design of highly active sgRNA is critical to accurate hit calling in such screens. Furthermore, while these screens have been broadly applied in mammalian cell lines and other model microbes, expansion to non-conventional organisms have been limited, in part due to the inability to accurately predict and design highly active sgRNA. Chapter 4 addresses this issue with the design of an experimental computation approach to sgRNA design that is specific to an organism of choice, in this case *Y. lipolytica.* CRISPR screens in the absence of the dominant DNA repair mechanism in this yeast (NHEJ) was used to generate guide activity profiles for both Cas9 and Cas12a. These, in addition to epigenetic data like nucleosome occupancy, served as input to design a deep learning sgRNA activity prediction algorithm called DeepGuide. Finally, DeepGuide ability to predict highly active guides for Cas9 and Cas12a was also independently validated on a subset of genes.

As discussed, a critical challenge in accurately assessing screening outcomes is accounting for the variability in guide activity. Poorly active guides targeting genes essential to screening conditions obscure the growth defects that are expected from disrupting them. While chapter 4 attempts to address this issue at the outset from the perspective of design, chapter 5 develops an end-to-end pipeline that identifies essential genes in pooled CRISPR screens for an existing library. It does so, by using experimentally determined cutting efficiencies for each guide in the library to provide an activity correction to the screening outcomes, thus accurately determining fitness effect of gene disruptions. Furthermore, a CRISPR Cas9 screen was utilized to investigate and discover known and novel genes that conferred tolerance to high salt and low pH conditions in *Y. lipolytica*. Finally, the outcomes of the preliminary Cas9 screen as well as DeepGuide activity predictions are used to design a smaller, optimized library, that is capable of accurate essential gene determination with less than half the size of the original library. Finally in chapter 6, the results presented in this dissertation are summarized and the broad conclusions and impact within the field is discussed, and possible routes forward are presented.

1.3 References

1. Richardson, B. From a fossil-fuel to a biobased economy: The politics of industrial biotechnology. Environ. Plann. C Gov. Policy 30, 282–296 (2012).

2. Aguilar, A., Wohlgemuth, R. & Twardowski, T. Perspectives on bioeconomy. N. Biotechnol. 40, 181–184 (2018).

3. Nielsen, J., Larsson, C., van Maris, A. & Pronk, J. Metabolic engineering of yeast for production of fuels and chemicals. Curr. Opin. Biotechnol. 24, 398–404 (2013).

4. Guo, M. & Song, W. The growing U.S. bioeconomy: Drivers, development and constraints. N. Biotechnol. 49, 48–57 (2019).

5. Carlson, R. Estimating the biotech sector's contribution to the US economy. Nat. Biotechnol. 34, 247–255 (2016).

6. Ledesma-Amaro, R., Dulermo, T. & Nicaud, J. M. Engineering Yarrowia lipolytica to produce biodiesel from raw starch. Biotechnol. Biofuels 8, 148 (2015).

7. Ledesma-Amaro, R. Microbial oils: A customizable feedstock through metabolic engineering. Eur. J. Lipid Sci. Technol. 117, 141–144 (2015).

8. Cho, J. S., Kim, G. B., Eun, H., Moon, C. W. & Lee, S. Y. Designing microbial cell factories for the production of chemicals. JACS Au 2, 1781–1799 (2022).

9. Lee, S. Y., Lee, D.-Y. & Kim, T. Y. Systems biotechnology for strain improvement. Trends Biotechnol. 23, 349–358 (2005).

10. Lee, S. Y. & Kim, H. U. Systems strategies for developing industrial microbial strains. Nat. Biotechnol. 33, 1061–1072 (2015).

11. Choi, K. R. et al. Systems metabolic engineering strategies: Integrating systems and synthetic biology with metabolic engineering. Trends Biotechnol. 37, 817–837 (2019).

12. Ko, Y.-S. et al. Tools and strategies of systems metabolic engineering for the development of microbial cell factories for chemical production. Chem. Soc. Rev. 49, 4615–4636 (2020).

13. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337, 816–821 (2012).

14. Caruthers, M. H. A brief review of DNA and RNA chemical synthesis. Biochem. Soc. Trans. 39, 575–580 (2011).

15. Thorwall, S., Schwartz, C., Chartron, J. W. & Wheeldon, I. Stress-tolerant non-conventional microbes enable next-generation chemical biosynthesis. Nat. Chem. Biol. 16, 113–121 (2020).

16. Renewable Fuels Association. Annual Ethanol Production. Renewable Fuels Association https://ethanolrfa.org/markets-and-statistics/annual-ethanol-production.

17. Löbs, A.-K., Schwartz, C. & Wheeldon, I. Genome and metabolic engineering in non-conventional yeasts: Current advances and applications. Synth Syst Biotechnol 2, 198–207 (2017).

18. Qiu, Z. & Jiang, R. Improving Saccharomyces cerevisiae ethanol production and tolerance via RNA polymerase II subunit Rpb7. Biotechnol. Biofuels 10, 125 (2017).

19. Kirimura, K. & Yoshioka, I. Citric Acid. in Comprehensive Biotechnology (ed. Moo-Young, M.) 158–165 (Elsevier, 2019).

20. García-Estrada, C., Martín, J. F., Cueto, L. & Barreiro, C. Omics approaches applied to Penicillium chrysogenum and penicillin production: Revealing the secrets of improved productivity. Genes (Basel) 11, 712 (2020).

21. Hong, K.-K. & Nielsen, J. Metabolic engineering of Saccharomyces cerevisiae: a key cell factory platform for future biorefineries. Cell. Mol. Life Sci. 69, 2671–2690 (2012).

22. Costa, D. A. et al. Physiological characterization of thermotolerant yeast for cellulosic ethanol production. Appl. Microbiol. Biotechnol. 98, 3829–3840 (2014).

23. Groeneveld, P., Stouthamer, A. H. & Westerhoff, H. V. Super life--how and why "cell selection" leads to the fastest-growing eukaryote. FEBS J. 276, 254–270 (2009).

24. Löser, C., Urit, T., Stukert, A. & Bley, T. Formation of ethyl acetate from whey by Kluyveromyces marxianus on a pilot scale. J. Biotechnol. 163, 17–23 (2013).

25. Shapouri, H. & Gallagher, P. Ethanol Cost-of-Production Survey. Agricultural Economics Report Number 841, (2002).

26. Breuer, U. & Harms, H. Debaryomyces hansenii--an extremophilic yeast with biotechnological potential. Yeast 23, 415–437 (2006).

27. Andreishcheva, E. N. et al. Adaptation to salt stress in a salt-tolerant strain of the yeast Yarrowia lipolytica. Biochemistry 64, 1061–1067 (1999).

28. Kurylenko, O. O. et al. Metabolic engineering and classical selection of the methylotrophic thermotolerant yeast Hansenula polymorpha for improvement of high-temperature xylose alcoholic fermentation. Microb. Cell Fact. 13, 122 (2014).

29. Shaw, A. J. et al. Metabolic engineering of a thermophilic bacterium to produce ethanol at high yield. Proc. Natl. Acad. Sci. U. S. A. 105, 13769–13774 (2008).

30. Zhang, B. et al. Improving ethanol and xylitol fermentation at elevated temperature through substitution of xylose reductase in Kluyveromyces marxianus. J. Ind. Microbiol. Biotechnol. 40, 305–316 (2013).

31. Ryabova, O., Chmil, O. & Sibirny, A. Xylose and cellobiose fermentation to ethanol by the thermotolerant methylotrophic yeast. FEMS Yeast Res. 4, 157–164 (2003).

32. Jones, D. T. & Woods, D. R. Acetone-butanol fermentation revisited. Microbiol. Rev. 50, 484–524 (1986).

33. Ramos, J. L., Duque, E., Huertas, M. J. & Haïdour, A. Isolation and expansion of the catabolic potential of a Pseudomonas putida strain able to grow in the presence of high concentrations of aromatic hydrocarbons. J. Bacteriol. 177, 3911–3916 (1995).

34. Matsushika, A., Negi, K., Suzuki, T., Goshima, T. & Hoshino, T. Identification and characterization of a novel Issatchenkia orientalis GPI-anchored protein, IoGas1, required for resistance to low pH and salt stress. PLoS One 11, e0161888 (2016).

35. Nishimasu, H. et al. Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell 156, 935–949 (2014).

36. Zetsche, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell 163, 759–771 (2015).

37. Li, B. et al. CRISPR-Cas12a possesses unconventional DNase activity that can be inactivated by synthetic oligonucleotides. Mol. Ther. Nucleic Acids 19, 1043–1052 (2020).

38. Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A. & Charpentier, E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. Nature 532, 517–521 (2016).

39. Port, F., Starostecka, M. & Boutros, M. Multiplexed conditional genome editing with Cas12a in Drosophila. Proc. Natl. Acad. Sci. U. S. A. 117, 22890–22899 (2020).

40. Campa, C. C., Weisbach, N. R., Santinha, A. J., Incarnato, D. & Platt, R. J. Multiplexed genome engineering by Cas12a and CRISPR arrays encoded on single transcripts. Nat. Methods 16, 887–893 (2019).

41. Ao, X. et al. A multiplex genome editing method for Escherichia coli based on CRISPR-Cas12a. Front. Microbiol. 9, 2307 (2018).

42. McCarty, N. S., Graham, A. E., Studená, L. & Ledesma-Amaro, R. Multiplexed CRISPR technologies for gene editing and transcriptional regulation. Nat. Commun. 11, 1281 (2020).

43. Ramesh, A., Ong, T., Garcia, J. A., Adams, J. & Wheeldon, I. Guide RNA engineering enables dual purpose CRISPR-Cpf1 for simultaneous gene editing and gene regulation in Yarrowia lipolytica. ACS Synth. Biol. 9, 967–971 (2020).

44. DiCarlo, J. E. et al. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res. 41, 4336–4343 (2013).

45. Gao, S. et al. Multiplex gene editing of the Yarrowia lipolytica genome using the CRISPR-Cas9 system. J. Ind. Microbiol. Biotechnol. 43, 1085–1093 (2016).

46. Weninger, A., Hatzl, A.-M., Schmid, C., Vogl, T. & Glieder, A. Combinatorial optimization of CRISPR/Cas9 expression enables precision genome engineering in the methylotrophic yeast Pichia pastoris. J. Biotechnol. 235, 139–149 (2016).

47. Finnigan, G. C. & Thorner, J. Complex in vivo Ligation Using Homologous Recombination and High-efficiency Plasmid Rescue from Saccharomyces cerevisiae. Bio Protoc. 5, (2015).

48. Yellman, C. M. Precise replacement of Saccharomyces cerevisiae proteasome genes with human orthologs by an integrative targeting method. G3 (Bethesda) 10, 3189–3200 (2020).

49. Horwitz, A. A. et al. Efficient multiplexed integration of synergistic alleles and metabolic pathways in yeasts via CRISPR-Cas. Cell Syst. 1, 88–96 (2015).

50. Löbs, A.-K., Engel, R., Schwartz, C., Flores, A. & Wheeldon, I. CRISPR-Cas9-enabled genetic disruptions for understanding ethanol and ethyl acetate biosynthesis in Kluyveromyces marxianus. Biotechnol. Biofuels 10, 164 (2017).

51. Cao, M. et al. Centromeric DNA facilitates nonconventional yeast genetic engineering. ACS Synth. Biol. 6, 1545–1553 (2017).

52. Schwartz, C. M., Hussain, M. S., Blenner, M. & Wheeldon, I. Synthetic RNA polymerase III promoters facilitate high-efficiency CRISPR-Cas9-mediated genome editing in Yarrowia lipolytica. ACS Synth. Biol. 5, 356–359 (2016).

53. Numamoto, M., Maekawa, H. & Kaneko, Y. Efficient genome editing by CRISPR/Cas9 with a tRNA-sgRNA fusion in the methylotrophic yeast Ogataea polymorpha. J. Biosci. Bioeng. 124, 487–492 (2017).

54. Qi, L. S. et al. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. Cell 184, 844 (2021).

55. Schwartz, C., Frogue, K., Ramesh, A., Misa, J. & Wheeldon, I. CRISPRi repression of nonhomologous end-joining for enhanced genome engineering via homologous recombination in Yarrowia lipolytica. Biotechnol. Bioeng. 114, 2896–2906 (2017).

56. Gilbert, L. A. et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell 154, 442–451 (2013).

57. Bock, C. et al. High-content CRISPR screening. Nature Reviews Methods Primers 2, 1–23 (2022).

58. Peters, J. M. et al. A comprehensive, CRISPR-based functional analysis of essential genes in bacteria. Cell 165, 1493–1506 (2016).

59. Liu, X. et al. High-throughput CRISPRi phenotyping identifies new essential genes in Streptococcus pneumoniae. Mol. Syst. Biol. 13, 931 (2017).

60. Yao, L. et al. Pooled CRISPRi screening of the cyanobacterium Synechocystis sp PCC 6803 for enhanced industrial phenotypes. Nat. Commun. 11, 1666 (2020).

61. Schwartz, C. et al. Validating genome-wide CRISPR-Cas9 function improves screening in the oleaginous yeast Yarrowia lipolytica. Metab. Eng. 55, 102–110 (2019).

62. Lian, J., Schultz, C., Cao, M., HamediRad, M. & Zhao, H. Multi-functional genome-wide CRISPR system for high throughput genotype-phenotype mapping. Nat. Commun. 10, 5794 (2019).

63. Coradetti, S. T. et al. Functional genomics of lipid metabolism in the oleaginous yeast Rhodosporidium toruloides. Elife 7, (2018).

64. Abt, T. D., Souffriau, B., Foulquié-Moreno, M. R., Duitama, J. & Thevelein, J. M. Genomic saturation mutagenesis and polygenic analysis identify novel yeast genes affecting ethyl acetate production, a non-selectable polygenic trait. Microb. Cell 3, 159–175 (2016).

65. Zhu, J. et al. Genome-Wide Determination of Gene Essentiality by Transposon Insertion Sequencing in Yeast Pichia pastoris. Sci. Rep. 8, 10223 (2018).

66. Yang, N., Wang, R. & Zhao, Y. Revolutionize genetic studies and crop improvement with high-throughput and genome-scale CRISPR/Cas9 gene editing technology. Mol. Plant 10, 1141–1143 (2017).

67. Egermeier, M., Russmayer, H., Sauer, M. & Marx, H. Metabolic Flexibility of Yarrowia lipolytica Growing on Glycerol. Front. Microbiol. 8, 49 (2017).

68. Magdouli, S., Guedri, T., Tarek, R., Brar, S. K. & Blais, J. F. Valorization of raw glycerol and crustacean waste into value added products by Yarrowia lipolytica. Bioresour. Technol. 243, 57–68 (2017).

69. Nambou, K. et al. Designing of a "cheap to run" fermentation platform for an enhanced production of single cell oil from Yarrowia lipolytica DSM3286 as a potential feedstock for biodiesel. Bioresour. Technol. 173, 324–333 (2014).

70. Thevenieau, F. et al. Uptake and assimilation of hydrophobic substrates by the oleaginous yeast Yarrowia lipolytica. in Handbook of Hydrocarbon and Lipid Microbiology 1513–1527 (Springer Berlin Heidelberg, 2010).

71. Narisetty, V. et al. Development of hypertolerant strain of Yarrowia lipolytica accumulating succinic acid using high levels of acetate. ACS Sustain. Chem. Eng. 10, 10858–10869 (2022).

72. Yaguchi, A., Spagnuolo, M. & Blenner, M. Engineering yeast for utilization of alternative feedstocks. Curr. Opin. Biotechnol. 53, 122–129 (2018).

73. Qiao, K., Wasylenko, T. M., Zhou, K., Xu, P. & Stephanopoulos, G. Lipid production in Yarrowia lipolytica is maximized by engineering cytosolic redox metabolism. Nat. Biotechnol. 35, 173–177 (2017).

74. Magnan, C. et al. Sequence Assembly of Yarrowia lipolytica Strain W29/CLIB89 Shows Transposable Element Diversity. PLoS One 11, e0162363 (2016).

75. Liu, L. & Alper, H. S. Draft genome sequence of the oleaginous yeast Yarrowia lipolytica PO1f, a commonly used metabolic engineering host. Genome Announc. 2, (2014).

76. Dourou, M., Aggeli, D., Papanikolaou, S. & Aggelis, G. Critical steps in carbon metabolism affecting lipid accumulation and their regulation in oleaginous microorganisms. Appl. Microbiol. Biotechnol. 102, 2509–2523 (2018).

77. Beopoulos, A. et al. Yarrowia lipolytica as a model for bio-oil production. Prog. Lipid Res. 48, 375–387 (2009).

78. Xu, P., Qiao, K., Ahn, W. S. & Stephanopoulos, G. Engineering Yarrowia lipolytica as a platform for synthesis of drop-in transportation fuels and oleochemicals. Proc. Natl. Acad. Sci. U. S. A. 113, 10848–10853 (2016).

79. Soong, Y.-H. V. et al. Microbial synthesis of wax esters. Metab. Eng. 67, 428–442 (2021).

80. Kamzolova, S. V. & Morgunov, I. G. Metabolic peculiarities of the citric acid overproduction from glucose in yeasts Yarrowia lipolytica. Bioresour. Technol. 243, 433–440 (2017).

81. Rzechonek, D. A., Dobrowolski, A., Rymowicz, W. & Mirończuk, A. M. Aseptic production of citric and isocitric acid from crude glycerol by genetically modified Yarrowia lipolytica. Bioresour. Technol. 271, 340–344 (2019).

82. Lei, Q., Zeng, W., Zhou, J. & Du, G. Efficient separation of α-ketoglutarate from Yarrowia lipolytica WSH-Z06 culture broth by converting pyruvate to l-tyrosine. Bioresour. Technol. 292, 121897 (2019).

83. Schwartz, C., Frogue, K., Misa, J. & Wheeldon, I. Host and Pathway Engineering for Enhanced Lycopene Biosynthesis in Yarrowia lipolytica. Front. Microbiol. 8, 2233 (2017).

84. Larroude, M. et al. A synthetic biology approach to transform Yarrowia lipolytica into a competitive biotechnological producer of β-carotene. Biotechnol. Bioeng. 115, 464–472 (2018).

85. Kildegaard, K. R. et al. Engineering of Yarrowia lipolytica for production of astaxanthin. Synth Syst Biotechnol 2, 287–294 (2017).

86. Liu, X. et al. Oil crop wastes as substrate candidates for enhancing erythritol production by modified Yarrowia lipolytica via one-step solid state fermentation. Bioresour. Technol. 294, 122194 (2019).

87. Blazeck, J. et al. Harnessing Yarrowia lipolytica lipogenesis to create a platform for lipid and biofuel production. Nat. Commun. 5, 3131 (2014).

88. Tai, M. & Stephanopoulos, G. Engineering the push and pull of lipid biosynthesis in oleaginous yeast Yarrowia lipolytica for biofuel production. Metab. Eng. 15, 1–9 (2013).

89. Bhutada, G. et al. Sugar versus fat: elimination of glycogen storage improves lipid accumulation in Yarrowia lipolytica. FEMS Yeast Res. 17, (2017).

90. Schwartz, C., Shabbir-Hussain, M., Frogue, K., Blenner, M. & Wheeldon, I. Standardized Markerless Gene Integration for Pathway Engineering in Yarrowia lipolytica. ACS Synth. Biol. 6, 402–409 (2017).

91. Schwartz, C., Curtis, N., Löbs, A.-K. & Wheeldon, I. Multiplexed CRISPR activation of cryptic sugar metabolism enables Yarrowia lipolytica growth on cellobiose. Biotechnol. J. 1700584 (2018).

92. Yang, Z., Edwards, H. & Xu, P. CRISPR-Cas12a/Cpf1-assisted precise, efficient and multiplexed genome-editing in Yarrowia lipolytica. Metab. Eng. Commun. 10, e00112 (2020).

Chapter 2: Guide RNA engineering enables dual purpose CRISPR-Cpf1 for simultaneous gene editing and gene regulation in *Yarrowia lipolytica*

2.1 Abstract

Yarrowia lipolytica has fast become a biotechnologically significant yeast for its ability to accumulate lipids to high levels. While there exists a suite of synthetic biology tools for genetic engineering in this yeast, there is a need for multipurposed tools for rapid strain generation. Here, we describe a dual purpose CRISPR-Cpf1 system that is capable of simultaneous gene disruption and gene regulation. Truncating guide RNA spacer length to 16 nt inhibited nuclease activity but not binding to the target loci, enabling gene activation and repression with Cpf1-fused transcriptional regulators. Gene repression was demonstrated using a Cpf1-Mxi1 fusion achieving a 7-fold reduction in mRNA, while CRISPR-activation with Cpf1-VPR increased hrGFP expression by 10-fold. High efficiency disruptions were achieved with gRNAs 23-25 bp in length, and efficiency and repression levels were maintained with multiplexed expression of truncated and full-length gRNAs. The developed CRISPR-Cpf1 system should prove useful in metabolic engineering, genome wide screening and functional genomics studies.

This chapter previously appeared as a Technical Note in *ACS synthetic biology*. The original citation is as follows: Ramesh, A., Ong, T., Garcia, J. A., Adams, J., & Wheeldon, I. (2020). Guide RNA engineering enables dual purpose CRISPR-Cpf1 for simultaneous gene editing and gene regulation in *Yarrowia lipolytica*. *ACS Synthetic Biology*, 9(4), 967-971.

2.2 Introduction

The non-conventional dimorphic yeast *Yarrowia lipolytica* has attracted attention as an industrially-relevant host due to its ability to utilize hydrocarbons and other non-sugars feedstocks as carbon sources for the production of high titers of intracellular lipids. Exploiting these phenotypes, metabolic engineers have designed strains that accumulate lipids to over 90% of yeast dry cell weight and titers as high as 85 g/L [1-2]. Modified fatty acid and lipid biosynthesis pathways have also been designed to produce commodity and high value chemicals such as long chain dicarboxylic acids, omega-3 fatty acids, and carotenoids among others [3-7].

Y. lipolytica's maturation as a host for chemical biosynthesis is in part due to new genetic engineering tools. CRISPR-Cas9 genome editing has played a large part in accelerating metabolic engineering efforts in this and other microbes [8-12]. Targeted genome editing in Yarrowia and other non-conventional yeast is challenging because DNA repair is dominated by non-homologous end joining, preventing the use of common synthetic biology tools that depend on the high capacity of *Saccharomyces cerevisiae* to perform homologous recombination [13-16]. CRISPR Cas9-based gene regulation and editing have helped mitigate this problem, but multiplexed and multi-functional synthetic biology tools for rapid strain engineering in Yarrowia are still needed. Cpf1, a family of Cas12a bacterial endonucleases, targets to genomic loci in a similar manner to Cas9 but has the advantage of processing its own CRISPR-RNA arrays [17]. The ability to mature its own guides RNAs (gRNAs) from a single transcript can be leveraged for easy multiplexing [18]. Cpf1 also benefits from a T-rich PAM sequence (TTTV) that does not overlap with Cas9 function,

and, unlike Cas9, does not require a tracrRNA sequence, which shortens gRNA expression cassettes [19].

Here, we demonstrate a dual function CRISPR-Cpf1 technology that simultaneously disrupts a gene target and regulates expression at other genomic loci. Length studies of Cpf1 gRNAs show that endonuclease function is lost with spacer sequences of 16 or less nucleotides (nt). We use this effect to control Cpf1 function by expressing guides of different lengths.

2.3 Results and Discussion

We first screened a series of Cpf1 orthologous from *Acidaminococcus spp.* BV3L6 (AsCpf1), *Lachnospiraceae bacterium* ND2006 (LbCpf1) and *Francisella novicida* U112 (FnCpf1). A single plasmid system containing both the Cpf1 and gRNA expression cassettes was used for gene disruption (Figure 2.1A). LbCpf1 showed the highest disruption efficiency in the preliminary screen (22 ± 5%; Figure S2.1) and was used for all subsequent experiments.

Three genes, MGA1, CAN1 and URA3, whose disruption produces an easily observed phenotype, were used to demonstrate and optimize multiplexed functionality. MGA1 knockout has been implicated in the suppression of pseudohyphal growth in yeast and null mutants are easily identifiable by a smooth surface colony that is distinct from the wild type rough morphology[20]. CAN1 null mutants are resistant to L-canavanine, which is structurally similar to arginine and toxic to cell growth[21]. Finally, the URA3 gene which is

responsible for the de novo synthesis of pyrimidines was selected as null mutants are auxotrophic for uracil and resistant to the Ura3 catalyzed product of 5-FOA[22].

Five gRNAs were designed and tested for each of MGA1, CAN1, and URA3. The best gRNA for each gene achieved disruption efficiencies of 89.5 ± 2.5%, 90.0 ± 5.7% and 74.4 ± 10% for MGA1, CAN1 and URA3, respectively, after 4 days of outgrowth (Figures 2.1B and S2.2). Other gRNAs produced lower disruptions efficiencies, but all were successful in creating double stranded breaks in the genome. The observed sequence-dependence of gRNA on endonuclease activity has been demonstrated on a genome-wide scale in *Y. lipolytica* using Cas9[20]. The same study also shows that Cas9 activity is influenced by chromatin structure, specifically that the nucleosome occupancy can hinder cutting. We anticipate similar relationships with gRNA sequence and nucleosome occupancy with Cpf1.

The best LbCpf1 gRNAs for each of MGA1, CAN1, and URA3 were used in multiplexed format to generate dual and triple knockouts. ΔMGA1-ΔCAN1 dual knockouts were produced in 55 ± 11% of the observed colonies (30/60, 41/60, 28/60), while the ΔMGA1-ΔURA3 and ΔCAN1-ΔURA3 mutants were generated with 59 ± 6% and 60 ± 6% efficiency (34/60, 33/60, 40/60; 35/60, 40/60, 33/60; Figures 1C and S3). Creating the triple knockout in a single experiment was less efficient with disruption of all genes occurring only 44 ± 4% of the time (40/90, 43/90, 36/90). These results are on par with a recent study of AsCpf1 in Yarrowia[23].

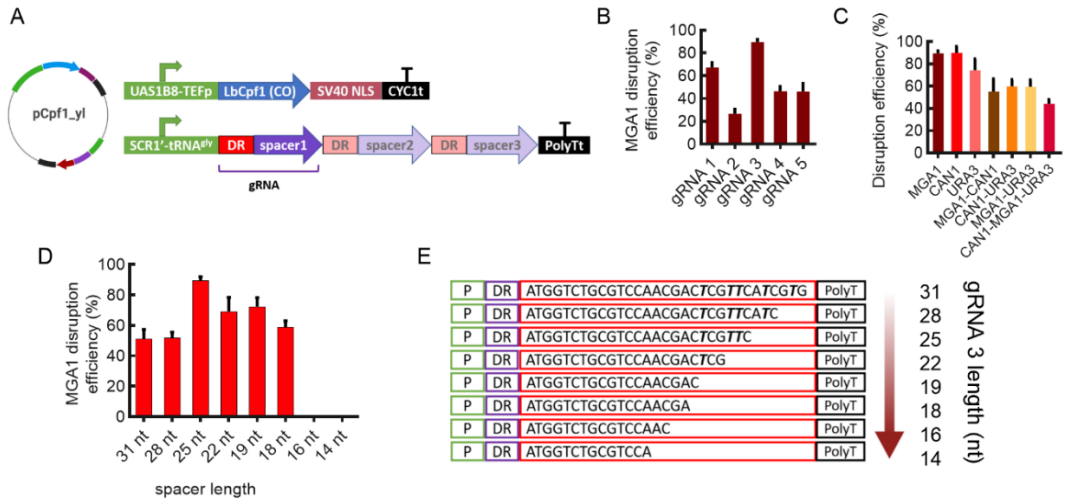

Figure 2.1. CRISPR-Cpf1 genome editing in *Yarrowia lipolytica*. (A) Schematic of the pCpf1_yl plasmid and expression cassettes for LbCpf1 and gRNA expression. Multiplexed cassettes are made by tiling direct repeat (DR) and spacer sequences. (B) Gene disruption efficiency for five different gRNAs targeting MGA1 in the PO1f strain of *Yarrowia lipolytica*. (C) Efficiency of double and triple disruptions of MGA1, CAN1 and URA3. (D) Effect of gRNA length on gene disruption efficiency, with guide sequences shown in (E). Thymine "T" nucleotides that are bolded and italicized indicate locations within each spacer where truncations were not made due to the presence of the polyT terminator. All *Y. lipolytica* transformants were grown in 2 mL of selective media in culture tubes at 30 °C. Data presented are mean and standard deviation of biological triplicates.

To characterize the effect of spacer length on LbCpf1 nuclease activity, the best gRNA for MGA1 and CAN1 were picked and the spacer length varied from 31 down to 14 nt. Expression of gRNAs with 23-25 nt spacers in the presence of active LbCpf1 resulted in the highest disruption efficiency for both MGA1 and CAN1. Endonuclease activity decreased in gRNAs longer than 25 and shorter than 23 (Figure 2.1D, E, S2.4, and S2.5). Most notably, cutting function sharply dropped with 14 and 16 nt spacers. None of the 90 screened colonies transformed with CRISPR plasmids expressing truncated spacers

showed phenotypic changes associated with MGA1 and CAN1 disruptions, and when 5 colonies were genotyped, none showed the presence of edits. This gRNA length-dependent effect is also seen with Cas9, which requires spacer of at least 16 nt to show detectable levels of gene editing in human cells[24].

Given the loss of Cpf1 endonuclease function at shorter spacer lengths and evidence that Cas9 binds target DNA but is not catalytically active with spacers 14 nt in length (see ref. 24), we explored the possibility of using active LbCpf1 with a fused repressor domain and truncated gRNA as a CRISPR interference (CRISPRi) system [25-26]. If shortened gRNAs can still form a ribonuclear complex with Cpf1 and bind to the genome loci complementary to the spacer sequence, then the system should function as a site-specific gene repressor. Swapping a repressor domain for an activation domain creates a gene activation tool.

CRISPRi studies have shown that transcriptional repression is effective when the endonuclease-repressor fusion is targeted within ~200 bp upstream of the transcription start site (TSS) [25-27]. We identified the putative TSS for CAN1 with the help of the YeasTSS online tool [28] and designed a series of five truncated gRNAs (t-gRNAs) with spacers 16 nt in length that span a short region surrounding the TSS (Figure 2.2A). In the case of CAN1, two putative TSS's were identified and targeted. A canavanine growth challenge revealed that the co-expression of LbCpf1 and t-gRNA2 enabled cell growth with cultures reaching an OD_{600} of 8.0 ± 1.3 after 48 hours, cell density significantly higher than the negative controls, one with no gRNA and a second with a scrambled gRNA that does not match a

loci within the genome (OD600 = 0.29 ± 0.03 and 0.23 ± 0.04, respectively; comparison $p<0.0001$, $n = 3$). In total, four out of five t-gRNAs (t-gRNA-1, -2, -4 and -5) showed a significant difference in growth to the negative controls ($p<0.05$; $n=3$). qPCR analysis of CAN1 transcript levels confirmed the repression effect, the two cultures that exhibited high resistance to canavanine (t-gRNA2 and -4) also had low levels of CAN1 mRNA with only 15.4 ± 3.1% and 33.5 ± 12.9% expression compared to the negative control. Importantly, sequencing of the region surrounding the targeted PAM sites revealed that endonuclease activity was not the cause of CAN1 downregulation (Figure S2.6). These results also compare well to a study of a deactivated FnCpf1-based CRISPRi study in Y. lipolytica [29].

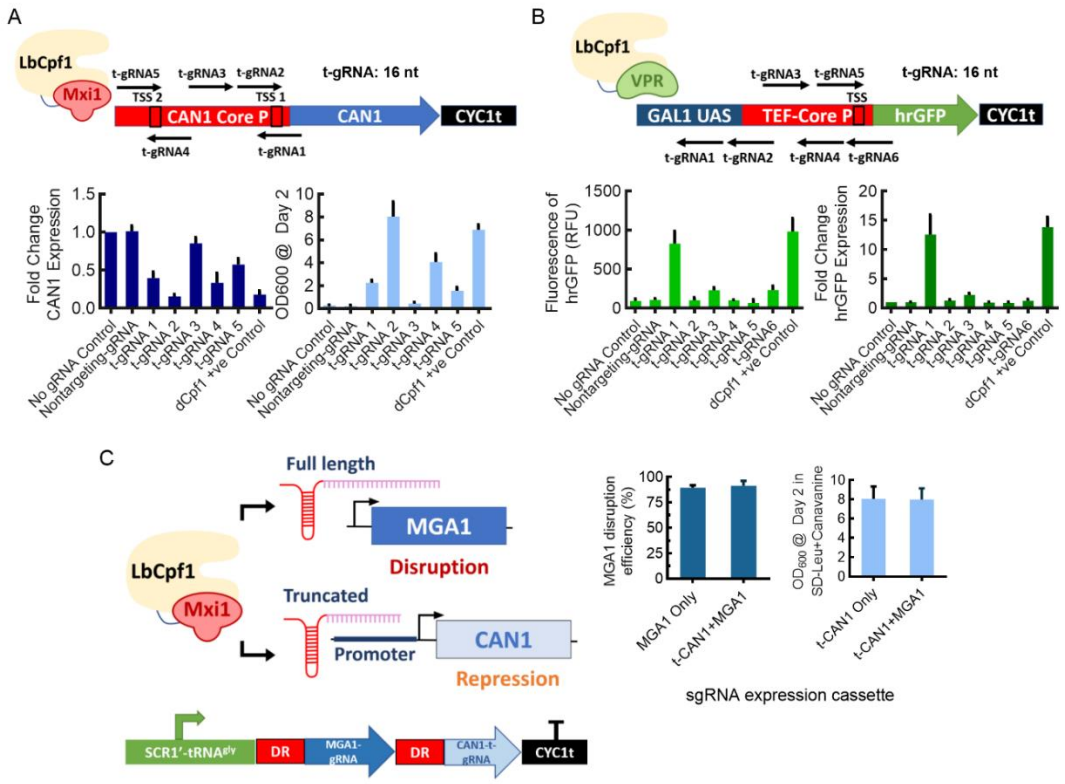

Figure 2.2. Truncated gRNAs enabled CRISPRa/i and dual functioning LbCpf1. (A) CRISPRi repression of CAN1 with truncated gRNAs and LbCpf1-MXI1. Repression of CAN1 with t-gRNA1, -2, -4 and -5 enables growth in a canavanine challenge assay. qPCR confirms reduced CAN1 mRNA levels correspond with increased growth. (B) CRISPRa activation of hrGFP with truncated gRNAs and LbCpf1-VPR. hrGFP expression, as measured by flow cytometry from a TEF core promoter with GAL1 UAS is low. Activation by CRISPRa with t-gRNA1 increases GFP fluorescence and hrGFP mRNA level. Basal autofluorescence was subtracted from all reported fluorescence values. Results in A and B are compared to negative controls with no gRNA and a nontargeting gRNA, as well as a positive control of CRISPRi/a enabled by deactivated Cpf1 (dCpf1) and full length gRNAs. (C) Simultaneous gene disruption and transcriptional repression using LbCpf1-Mxi1. A dual gRNA expression system producing t-gRNA2 for CAN1 and a gRNA for MGA1 disruption effectively repressed CAN1 while editing MGA1. All *Y. lipolytica* transformants were grown in 2 mL of selective media in culture tubes at 30 °C. Data presented are mean and standard deviation of biological triplicates.

Our previous CRISPR activation (CRISPRa) studies with deactivated Cas9 fused to the synthetic transcriptional activator VPR also revealed that function varies with distance from the TSS [30]. Again, we used a series of t-gRNAs that span a region upstream of the gene of interest, in this case an engineered GFP expression cassette integrated at the XPR2 locus of *Y. lipolytica* PO1f (Figure 2.2B). Six t-gRNAs were designed that span ~150 bp of the GAL1-TEF_{core} promoter that drives expression of the integrated cassette. CRISPR plasmids expressing one guide and LbCpf1 were transformed into PO1f and random colonies were selected for flow cytometry and qPCR analysis. A CRISPRa plasmid that expressed no gRNA, as well as one that expressed a non-targeting gRNA, were used as negative controls. One out of six sgRNAs (t-gRNA1) showed significant activation at nearly 10-fold above the negative controls. None of the other gRNAs showed any appreciable levels of activation. Sequencing the regions surrounding the targeted PAM site for the best performer, revealed no edits (Figure S2.6).

For both the CRISPRi and CRISPRa studies, we also performed positive control experiments using deactivated LbCpf1 (Cpf1 D832A; dCpf1). In these experiments, dLbCpf1 was co-expressed with the full-length t-gRNAs that showed the best result for activation and repression. Random colonies were subjected to canavanine toxicity challenge (for CRISPRi), flow cytometry (for CRISPRa) and qPCR analysis. These experiments showed that our developed CRISPRi/a system that uses active Cpf1 and truncated gRNAs performs just as well traditional technologies (Figures 2A and B).

Together, the length study data and CRISPRi/a demonstrations show that LbCpf1 endonuclease activity can be controlled through gRNA expression. This presents the opportunity to create arrays of guides that target different gene editing functions (disruption, activation, and repression) to sites throughout the genome. To this end, we designed a dual function CRISPR-Cpf1 system by simultaneously expressing a full-length spacer for one gene, MGA1, and a truncated 16 nt spacer for a second, CAN1. The dual expression system was successful. After 2 days of outgrowth in selective media, cultures were subjected to a canavanine toxicity challenge, phenotyped, and genotyped for MGA1 disruption. Dual expression did not affect Cpf1 and CRISPRi function; MGA1 disruption occurred at 92.4 ± 6.1% efficiency and growth in the toxicity challenge was equivalent to the control (Figure 2.2C). We also note that Mxi1 and VPR fusion to LbCpf1 had no effect on nuclease activity with full length gRNAs (Figure S2.6).

In studying the effect of Cpf1 gRNA length on endonuclease activity we identified a switch point in function. Spacers 16 nt in length bind to the target site but do not produce double stranded breaks. Spacers greater than 16 nt and up to 31 nt activate LbCpf1 activity. These results are consistent with analyses of Cpf1 crystal structures. Specifically, that the 5'-stem loop of the direct repeat is necessary and sufficient for the formation of a ribonuclear complex, and that the endonuclease domains interact with the genomic target at the 23rd and the 18th positions of the spacer [31-33]. Given this, we speculate that gRNA shorter than 18 nt are unable to activate endonuclease activity but maintain sufficient homology to attach the ribonuclear complex to the locus of interest. Here, we leveraged this effect to express Cpf1 CRISPR-RNA arrays with gRNAs of different lengths, along

with LbCpf1 fused to an activator for CRISPRa, or a repressor domain for CRISPRi, to enable multifunctional genome editing. Synthetic biology tools that enable rapid and multiplexed genome modifications are needed to overcome a bottleneck in non-conventional yeast strain engineering. The dual function CRISPR-Cpf1 system shown here adds to the tools needed to address this challenge.

2.4 Associated Content

2.4.1 Supporting Information

Methods; Cpf1 nucleotide sequences; yeasts strains, plasmids, and primers used in this study; initial screening LbCpf1 and FnCpf1 endonuclease activity in *Y. lipolytica*; MGA1, CAN1, and URA3 single and double disruptions and phenotypes; gRNA length study for the disruption of CAN1; sequence alignments of MGA1 and CAN1 targeted by various gRNA lengths; sequence alignments of CAN1 and MGA1 showing indels resulting from Cpf1 endonuclease activity; sequence alignments of CAN1 and hrGFP promoters targeted by truncated gRNAs; gene disruption efficiency effected LbCpf1 fusions with transcriptional regulators; method comparison to other CRISPR-Cpf1 tools in *Y. lipolytica*.

2.5 Author Information

2.5.1 Corresponding Author

*Email: iwheeldon@engr.ucr.edu

2.5.2 Notes

The authors declare no competing financial interests.

2.5.3 Author Contributions

AR and IW conceived the study, analyzed the data and wrote the paper. TO, JAG, JA, and AR conducted the experiments. All authors edited the manuscript.

2.5.4 Conflict of Interest

The authors declare that they have no conflicts of interest.

2.6 Acknowledgements

This study was supported by NSF-CBET 1706545 to IW and NSF-REU 1461297 to the UC-Riverside Center for Plant Cell Biology.

2.7 References

1. Qiao, K. J.; Wasylenko, T. M.; Zhou, K.; Xu, P.; Stephanopoulos, G., Lipid production in Yarrowia lipolytica is maximized by engineering cytosolic redox metabolism. *Nat. Biotechnol.* **2017,** *35* (2), 173-177.

2. Blazeck, J.; Hill, A.; Liu, L. Q.; Knight, R.; Miller, J.; Pan, A.; Otoupal, P.; Alper, H. S., Harnessing Yarrowia lipolytica lipogenesis to create a platform for lipid and biofuel production. *Nat. Commun.* **2014,** *5*.

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

4. Xue, Z. X.; Sharpe, P. L.; Hong, S. P.; Yadav, N. S.; Xie, D. M.; Short, D. R.; Damude, H. G.; Rupert, R. A.; Seip, J. E.; Wang, J.; Pollak, D. W.; Bostick, M. W.; Bosak, M. D.; Macool, D. J.; Hollerbach, D. H.; Zhang, H. X.; Arcilla, D. M.; Bledsoe, S. A.; Croker, K.; McCord, E. F.; Tyreus, B. D.; Jackson, E. N.; Zhu, Q., Production of omega-3 eicosapentaenoic acid by metabolic engineering of Yarrowia lipolytica. *Nat. Biotechnol.* **2013,** *31* (8), 734-+.

5. Blazeck, J.; Liu, L. Q.; Knight, R.; Alper, H. S., Heterologous production of pentane in the oleaginous yeast Yarrowia lipolytica. *J Biotechnol.* **2013,** *165* (3-4), 184-194.

6. Ledesma-Amaro, R.; Nicaud, J. M., Yarrowia lipolytica as a biotechnological chassis to produce usual and unusual fatty acids. *Prog Lipid Res.* **2016,** *61*, 40-50.

7. Markham, K. A.; Palmer, C. M.; Chwatko, M.; Wagner, J. M.; Murray, C.; Vazquez, S.; Swaminathan, A.; Chakravarty, I.; Lynd, N. A.; Alper, H. S., Rewiring Yarrowia lipolytica toward triacetic acid lactone for materials generation. *Proc. Natl. Acad. Sci. U. S. A.* **2018,** *115* (9), 2096-2101.

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

9. Schwartz, C.; Shabbir-Hussain, M.; Frogue, K.; Blenner, M.; Wheeldon, I., Standardized Markerless Gene Integration for Pathway Engineering in Yarrowia lipolytica. *ACS Synth. Biol.* **2017,** *6* (3), 402-409.

10. Schwartz, C. M.; Hussain, M. S.; Blenner, M.; Wheeldon, I., Synthetic RNA Polymerase III Promoters Facilitate High-Efficiency CRISPR-Cas9-Mediated Genome Editing in Yarrowia lipolytica. *ACS Synth. Biol.* **2016,** *5* (4), 356-359.

11. Cook, T. B.; Rand, J. M.; Nurani, W.; Courtney, D. K.; Liu, S. A.; Pfleger, B. F., Genetic tools for reliable gene expression and recombineering in Pseudomonas putida. *J. Ind. Microbiol. Biotechnol.* **2018,** *45* (7), 517-527.

12. Tran, V. G.; Cao, M.; Fatma, Z.; Song, X.; Zhao, H., Development of a CRISPR/Cas9-Based Tool for Gene Deletion in <em>Issatchenkia orientalis</em&gt. *mSphere* **2019,** *4* (3), e00345-19.

13. Löbs, A.-K.; Schwartz, C.; Wheeldon, I., Genome and metabolic engineering in non-conventional yeasts: Current advances and applications. *Synth. Syst. Biotechnol.* **2017,** *2* (3), 198-207.

14. Shao, Z. Y.; Zhao, H.; Zhao, H. M., DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Res.* **2009,** *37* (2).

15. Horwitz, Andrew A.; Walter, Jessica M.; Schubert, Max G.; Kung, Stephanie H.; Hawkins, K.; Platt, Darren M.; Hernday, Aaron D.; Mahatdejkul-Meadows, T.; Szeto, W.; Chandran, Sunil S.; Newman, Jack D., Efficient Multiplexed Integration of Synergistic Alleles and Metabolic Pathways in Yeasts via CRISPR-Cas. *Cell Syst.* **2015,** (1), 1-9.

16. Sadhu, M. J.; Bloom, J. S.; Day, L.; Siegel, J. J.; Kosuri, S.; Kruglyak, L., Highly parallel genome variant engineering with CRISPR-Cas9. *Nat. Genet.* **2018,** *50* (4), 510-+.

17. Fonfara, I.; Richter, H.; Bratovic, M.; Le Rhun, A.; Charpentier, E., The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **2016,** *532* (7600), 517-+.

18. Zetsche, B.; Heidenreich, M.; Mohanraju, P.; Fedorova, I.; Kneppers, J.; DeGennaro, E. M.; Winblad, N.; Choudhury, S. R.; Abudayyeh, O. O.; Gootenberg, J. S.; Wu, W. Y.; Scott, D. A.; Severinov, K.; van der Oost, J.; Zhang, F., Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. *Nat. Biotechnol.* **2017,** *35* (1), 31-34.

19. Zetsche, B.; Gootenberg, J. S.; Abudayyeh, O. O.; Slaymaker, I. M.; Makarova, K. S.; Essletzbichler, P.; Volz, S. E.; Joung, J.; van der Oost, J.; Regev, A.; Koonin, E. V.; Zhang, F., Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **2015,** *163* (3), 759-771.

20. 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. *Metab. Eng.* **2019,** *55*, 102-110.

21. Fantes, P. A.; Creanor, J., Canavanine resistance and the mechanism of arginine uptake in the fission yeast Schizosaccharomyces pombe. *Microbiology* **1984,** *130* (12), 3265-3273.

22. Boeke, J. D.; Trueheart, J.; Natsoulis, G.; Fink, G. R., [10] 5-Fluoroorotic acid as a selective agent in yeast molecular genetics. In *Methods Enzymol.*, Elsevier: 1987; Vol. 154, pp 164-175.

23. Yang, Z.; Edwards, H.; Xu, P., CRISPR-Cas12a/Cpf1-assisted precise, efficient and multiplexed genome-editing in Yarrowia lipolytica. *Metab. Eng. Commun.* **2020,** *10*, e00112.

24. Kiani, S.; Chavez, A.; Tuttle, M.; Hal, R. N.; Chari, R.; Ter-Ovanesyan, D.; Qian, J.; Pruitt, B. W.; Beal, J.; Vora, S.; Buchthal, J.; Kowal, E. J. K.; Ebrahimkhani, M. R.; Collins, J. J.; Weiss, R.; Church, G., Cas9 gRNA engineering for genome editing, activation and repression. *Nat. Methods* **2015,** *12* (11), 1051-1054.

25. Schwartz, C.; Frogue, K.; Ramesh, A.; Misa, J.; Wheeldon, I., CRISPRi repression of nonhomologous end-joining for enhanced genome engineering via homologous recombination in Yarrowia lipolytica. *Biotechnol. Bioeng.* **2017,** *114* (12), 2896-2906.

26. Lobs, A. K.; Schwartz, C.; Thorwall, S.; Wheeldon, I., Highly Multiplexed CRISPRi Repression of Respiratory Functions Enhances Mitochondrial Localized Ethyl Acetate Biosynthesis in Kluyveromyces marxianus. *ACS Synth. Biol.* **2018,** *7* (11), 2647-2655.

27. Deaner, M.; Alper, H. S., Systematic testing of enzyme perturbation sensitivities via graded dCas9 modulation in Saccharomyces cerevisiae. *Metab. Eng.* **2017,** *40*, 14-22.

28. McMillan, J.; Lu, Z.; Rodriguez, J. S.; Ahn, T.-H.; Lin, Z., YeasTSS: an integrative web database of yeast transcription start sites. *Database* **2019,** *2019*.

29. Zhang, J.-l.; Peng, Y.-Z.; Liu, D.; Liu, H.; Cao, Y.-X.; Li, B.-Z.; Li, C.; Yuan, Y.-J., Gene repression via multiplex gRNA strategy in Y. lipolytica. *Microb. Cell Fact.* **2018,** *17* (1), 62.

30. Schwartz, C.; Curtis, N.; Lobs, A. K.; Wheeldon, I., Multiplexed CRISPR Activation of Cryptic Sugar Metabolism Enables Yarrowia Lipolytica Growth on Cellobiose. *Biotechnol. J.* **2018,** *13* (9).

31. Li, B.; Zeng, C.; Dong, Y., Design and assessment of engineered CRISPR–Cpf1 and its use for genome editing. *Nat. Protoc.* **2018,** *13* (5), 899.

32. Dong, D.; Ren, K.; Qiu, X.; Zheng, J.; Guo, M.; Guan, X.; Liu, H.; Li, N.; Zhang, B.; Yang, D., The crystal structure of Cpf1 in complex with CRISPR RNA. *Nature* **2016,** *532* (7600), 522-526.

33. Yamano, T.; Nishimasu, H.; Zetsche, B.; Hirano, H.; Slaymaker, I. M.; Li, Y.; Fedorova, I.; Nakane, T.; Makarova, K. S.; Koonin, E. V., Crystal structure of Cpf1 in complex with guide RNA and target DNA. *Cell* **2016,** *165* (4), 949-962.

2.8 Supplementary Information

2.8.1 Methods

2.8.1.1 Strains, cultures, and transformations

The *Escherichia coli strain* TOP10 was used for the construction and propagation of all plasmids, and was cultured in Luria-Bertani broth with 100 mg/L ampicillin. *E. coli* was cultured at 37 °C in 14 mL polypropylene tubes, at 225 RPM. Plasmids were isolated from *E. coli* cultures using the Zymo Research Plasmid Miniprep Kit. *Y. lipolytica* PO1f (MatA, leu2-270, ura3-302, xpr2-322, axp-2), PO1f URA3::A08, and PO1f GAL1UAS-TEF-hrGFP::XPR2 were grown in YPD medium (2% Bacto peptone, 1% Bacto yeast extract, 2% glucose) or on YPD agar plates (2% agar). Strains transformed with a plasmid were grown in synthetic defined medium without leucine (SD-leu; 0.069% CSM-leu (Sunrise Science), 0.67% Difco yeast nitrogen base without amino acids, and 2% glucose) or on SD-leu agar plates (2% agar). *Y. lipolytica* stationary phase transformations were done using a modified PEG-LiAC protocol as described in a previous work [1-2]. All *Y. lipolytica* strains (Table S2.1) were cultured at 30 °C in 14 mL polypropylene tubes, at 225 RPM[3].

2.8.1.2 Plasmid design and cloning

LbCpf1 sequence was obtained from plasmid SQT1665 (Addgene Plasmid #78744), and then codon optimized for use in *Y. lipolytica*[4]. The amino acid sequence was used as input to Optimizer (http://genomes.urv.es/OPTIMIZER/) with the codon usage table of the CLIB122 strain of *Y. lipolytica*[5]. In a similar fashion, the FnCpf1

sequence was obtained from Addgene Plasmid #69976 and codon optimized for use in *Y. lipolytica* [6].

All enzymes for cloning purposes were purchased from New England Biolabs. Q5 DNA polymerase was used to perform PCR reactions and all Gibson Assembly reactions were done using the NEBuilder® HiFi DNA Assembly mix. PCR purifications were done using the Zymo DNA Clean and Concentrator kit. To generate the LbCpf1 CRISPR plasmid, pUAS1B8-TEF(136)-hrGFP was digested with BssHII and NheI. Digested backbone and LbCpf1 fragment were purified using the Zymo Gel DNA Extraction kit, and cloned using the T4 DNA ligase kit from NEB. The resulting plasmid was digested with AatII and the digestion product was then used as the backbone to clone in the gRNA expression cassette by Gibson Assembly[7]. Primers SCR_DR_F, SCR_DR_R_Lb and SCR_DR_R_Fn (see Table S2.2) were used to amplify the gRNA expression cassette from the previously generated pCRISPRyl plasmid [8]. The resulting CRISPR-Cpf1 cloning vectors contained a SCR1'-tRNA^{gly} PolIII promoter, a 20 nt LbCpf1 or FnCpf1 direct repeat, a SpeI cloning site to insert the gRNA, and a PolyT terminator. All gRNAs inserts were ordered as single strand primers with overlaps to enable Gibson Assembly with respective Cpf1 expression vector. To generate the FnCpf1 and LbCpf1 CRISPR plasmids targeting PEX10, 6 primers PEX_Sg1_Fn, PEX_Sg2_Fn, PEX_Sg3_Fn, PEX_Sg1_Lb, PEX_Sg2_Lb, and PEX_Sg3_Lb containing the 3 sgRNA targeting PEX10 were ordered and cloned into the FnCpf1 and LbCpf1 cloning vectors via Gibson Assembly. The LbCpf1 cloning vector (pCpf1_yl) was used for all further cloning.

To generate the single gene knockout CRISPR plasmids targeting MGA1, CAN1, and URA3, primers MGA_Sg1, MGA_Sg2, MGA_Sg3, MGA_Sg4, MGA_Sg5, CAN_Sg1, CAN_Sg2, CAN_Sg3, CAN_Sg4, CAN_Sg5, URA_Sg1, URA_Sg2, URA_Sg3, URA_Sg4, and URA_Sg5 encoding the respective gRNAs with overlaps to the pCpf1_yl backbone were ordered and cloned into the Cpf1 expression vector by Gibson Assembly. The MGA1-CAN1 dual knockout plasmid was cloned using the MGA_MC and CAN_MC primer set. Similarly, the CAN1-URA3 and MGA1-URA3 dual knockout plasmids were generated using the CAN_CU, URA_CU and MGA_MU, URA_MU primer sets. The sgRNA for the triple knockout plasmid was cloned using CMU_1 and CMU_2 primers. The plasmids containing varying lengths of the best performing gRNA for MGA1 and CAN1 were cloned in a similar manner. Primers MGA_31, MGA_28, MGA_22, MGA_19, MGA_18, MGA_16, and MGA_14 were used to generate plasmids targeting MGA1. Similarly, primers CAN_30, CAN_28, CAN_22, CAN_20, CAN_18, CAN_16, and CAN_14 were used make plasmids targeting CAN1 with gRNA ranging from 30 to 14 nt.

For the generation of the CRISPRi cloning vector, Mxi1_F and Mxi1_R primers were used to amplify the Mxi1 repression domain from the previously described pCRISPRi_yl plasmid [8]. pCpf1_yl was digested with NheI and the digestion product was used to clone in the Mxi1 domain using Gibson Assembly to generate pCpf1i_yl. The cloning vector for CRISPRa was generated by first digesting pCpf1_yl with NheI. The VPR activator was amplified from the previously described pCRISPRa_VPR_yl plasmid using primers VPR_F and VPR_R [2], and then assembled into the digested pCpf1yl vector

to generate pCpf1a_yl. The truncated gRNA targeting the CAN1 promoter region for repression were cloned into pCpf1i_yl using the primers CAN_Tsg1, CAN_Tsg2, CAN_Tsg3, CAN_Tsg4, and CAN_Tsg5. Similarly, the truncated sgRNA targeting upstream of hrGFP were cloned into pCpf1a_yl using primers GFP_Tsg1, GFP_Tsg2, GFP_Tsg3, GFP_Tsg4, GFP_Tsg5, and GFP_Tsg6.

2.8.1.3 Screening for gene disruption

To screen for MGA1 gene disruption, cultures with CRISPR plasmids growing in SD-Leu were diluted and plated in triplicate on YPD to obtain greater than 50 colonies on each plate. After 2 days of growth at 30 °C, the number of smooth colonies were then counted and expressed as a fraction of total colonies on the plate. For disruption of the CAN1 gene, cultures were similarly diluted and plated on YPD to obtain single colonies. Thirty colonies in triplicate were then randomly selected and streaked on SD media supplemented with 50 mg/L of L-canavanine. Colonies that grew on SD+canavanine were identified as positive for CAN1 disruption. To screen for URA3, cultures were similarly plated, and 30 colonies in triplicated were randomly selected and streaked on YPD+5FOA and SD-Ura. Growth on SD-Ura but not on YPD-5FOA indicated URA3 disruption. Confirmation of MGA1 and CAN1 disruptions were obtained by sequencing 8 randomly selected colonies.

Triplicates of 60 random colonies were screened when confirming dual knockouts, and triplicates of 90 random colonies were screened for triple knockouts. For example, a dual knockout of MGA1 and CAN1 was screened by plating cultures and

selecting 60 random colonies in triplicates before the colonies were grown enough to distinguish between smooth and rough morphologies. Then the selected colonies are streaked out on SD+canavanine. The colonies that show both smooth morphologies and growth on SD+canavanine were considered disrupted for both genes.

2.8.1.4 Design and selection of gRNA

For gene disruption, 23-25 nt gRNAs with a TTTV PAM sequence (V=A/G/C) were designed and checked for uniqueness by BLAST search against the *Y. lipolytica* PO1f genome. For CRISPRi repression of CAN1, we first identified putative transcription start sites (TSS) of CAN1 using the YeasTSS webtool (http://www.yeastss.org/)[9]. gRNAs were designed to target around the TSS and within 200 bp upstream of the TSS. For the CRISPRa activation of hrGFP, the gene was first expressed from a GAL4UAS-TEFmin promoter to attain minimal hrGFP expression. Subsequently, all sgRNA upstream of the start codon within the GAL4UAS-TEFmin region were designed and investigated.

2.8.1.5 RT-qPCR

Y. lipolytica transformants were grown to early stationary phase in SD-Leu (OD_{600} ~10) and subjected to RNA extraction. One-mL of a culture at an OD_{600} of 10 was spun down by centrifugation at 6,500g for 2 min. The Yeastar RNA isolation Kit from Zymo Research was then used to isolate total RNA. The resulting RNA was then subjected to DNaseI digestion for 45 min at 37°C to prevent genomic DNA

contamination. The DNA-free RNA was then purified using the RNA Clean and Concentrator-25 kit from Zymo Research. RNA concentration was quantified spectrophotometry to determine the presence of any genomic DNA contamination. 400 ng of purified RNA was used to generate cDNA with the iScript Reverse Transcription Supermix from Biorad and the remaining stored at -80 °C. The resulting cDNA was diluted 8-fold, and 2 µl was used in each RT-qPCR experiment with the SsoAdvanced Universal SYBR Green Supermix from Biorad and appropriate primers. All experiments were performed in biological triplicates and technical duplicates using 96-well plates on a Biorad CFX Connect Thermocycler. Primers for qPCR were designed on the IDT PrimerQuest tool according to specifications present in the SsoAdvanced SYBR Green Supermix manual. A primer efficiency curve was also generated and the primers were validated to have an efficiency of between 0.90 and 1.10 before use in the qPCR experiments. Finally, primers qCAN_F and qCAN_R were used to amplify CAN1 for qPCR, and qGFP_F and qGFP_R were used to amplify hrGFP. Relative expression levels and later fold change mRNA expression were determined by normalizing to the expression of a housekeeping gene (actin). Actin amplification was achieved using primers Act_F and Act_R.

2.8.1.6 Flow Cytometry

CRISPRa plasmids containing the sgRNA targeting the GAL1UAS-TEFcore sequence were transformed into *Y. lipolytica* and plated on SD-Leu media. A single colony from each plate was used to inoculate 2 mL SD-Leu liquid cultures in 14 mL

polypropylene tubes. Stationary phase cells (OD_{600} ~10) were spun down at 6500g for 2 min, washed twice with 1X phosphate buffered saline (PBS) solution and resuspended in 200 µL water. The BD accuri C6 flow cytometer was used for data collection and analysis. A control strain not expressing hrGFP was used to identify basal autofluorescence prior to collecting data for the experimental samples. The control strain was transformed with a vector containing the Leu cassette and grown in SD-Leu. Basal autofluorescence was identified to be ~300 AU which was subtracted from all reported fluorescence values. For all samples, the population of healthy *Y. lipolytica* cells were gated in FSC-SSC plot and 10,000 events were collected in this gate. All experiments were performed in biological triplicates.

2.8.1.7 Canavanine growth challenge

To assess the phenotypic effect of the CRISPRi repression of CAN1, a canavanine growth challenge was performed. The CRISPRi plasmids containing gRNA targeting CAN1 were transformed into *Y. lipolytica* and plated on SD-Leu media. A single colony from each plate was used to inoculate 2 mL SD-Leu liquid cultures in 14 mL culture tubes. Stationary phase cells (OD_{600} ~10) were subject to a canavanine challenge in 2 mL SD-Leu, 50 mg/L L-canavanine with an initial OD_{600} of 0.1. Cell density was measured after 48 hours. *Y. lipolytica* transformed with a CRISPRi plasmid containing no sgRNA was used as a negative control for the experiment.

2.8.2 Nucleotide Sequence Information

2.8.2.1 Nucleotide sequence of *Y. lipolytica* codon optimized LbCpf1-SV40

Sequence of *Y. lipolytica* codon optimized LbCpf1.

ATGTCTAAGCTGGAGAAGTTCACCAACTGCTACTCTCTGTCTAAGACCCTGCGATTCAA
GGCCATCCCCGTGGGCAAGACCCAGGAGAACATCGACAACAAGCGACTGCTGGTGGAGG
ACGAGAAGCGAGCCGAGGACTACAAGGGCGTGAAGAAGCTGCTGGACCGATACTACCTG
TCTTTCATCAACGACGTGCTGCACTCTATCAAGCTGAAGAACCTGAACAACTACATCTC
TCTGTTCCGAAAGAAGACCCGAACCGAGAAGGAGAACAAGGAGCTGGAGAACCTGGAGA
TCAACCTGCGAAAGGAGATCGCCAAGGCCTTCAAGGGCAACGAGGGCTACAAGTCTCTG
TTCAAGAAGGACATCATCGAGACCATCCTGCCCGAGTTCCTGGACGACAAGGACGAGAT
CGCCCTGGTGAACTCTTTCAACGGCTTCACCACCGCCTTCACCGGCTTCTTCGACAACC
GAGAGAACATGTTCTCTGAGGAGGCCAAGTCTACCTCTATCGCCTTCCGATGCATCAAC
GAGAACCTGACCCGATACATCTCTAACATGGACATCTTCGAGAAGGTGGACGCCATCTT
CGACAAGCACGAGGTGCAGGAGATCAAGGAGAAGATCCTGAACTCTGACTACGACGTGG
AGGACTTCTTCGAGGGCGAGTTCTTCAACTTCGTGCTGACCCAGGAGGGCATCGACGTG
TACAACGCCATCATCGGCGGCTTCGTGACCGAGTCTGGCGAGAAGATCAAGGGCCTGAA
CGAGTACATCAACCTGTACAACCAGAAGACCAAGCAGAAGCTGCCCAAGTTCAAGCCCC
TGTACAAGCAGGTGCTGTCTGACCGAGAGTCTCTGTCGTTCTACGGCGAGGGCTACACC
TCTGACGAGGAGGTGCTGGAGGTGTTCCGAAACACCCTGAACAAGAACTCTGAGATCTT
CTCTTCTATCAAGAAGCTGGAGAAGCTGTTCAAGAACTTCGACGAGTACTCTTCTGCCG
GCATCTTCGTGAAGAACGGCCCCGCCATCTCTACCATCTCTAAGGACATCTTCGGCGAG
TGGAACGTGATCCGAGACAAGTGGAACGCCGAGTACGACGACATCCACCTGAAGAAGAA
GGCCGTGGTGACCGAGAAGTACGAGGACGACCGACGAAAGTCTTTCAAGAAGATCGGCT
CTTTCTCTCTGGAGCAGCTGCAGGAGTACGCCGACGCCGACCTGTCTGTGGTGGAGAAG
CTGAAGGAGATCATCATTCAGAAGGTGGACGAGATCTACAAGGTGTACGGCTCTTCCGA
GAAGCTGTTTGACGCTGACTTCGTGCTGGAGAAGTCTCTGAAGAAGAACGACGCCGTGG
TGGCCATCATGAAGGACCTGCTGGACTCTGTGAAGTCTTTCGAGAACTACATCAAGGCC
TTCTTCGGCGAGGGCAAGGAGACCAACCGAGACGAGTCTTTCTACGGCGACTTCGTGCT
GGCCTACGACATCCTGCTGAAGGTGGACCACATCTACGACGCCATCCGAAACTACGTGA
CCCAGAAGCCCTACTCTAAGGACAAGTTCAAGCTGTACTTCCAGAACCCCCAGTTCATG
GGCGGCTGGGACAAGGACAAGGAGACCGACTACCGAGCCACCATCCTGCGATACGGCTC
TAAGTACTACCTGGCCATCATGGACAAGAAGTACGCCAAGTGCCTGCAGAAGATCGACA
AGGACGACGTGAACGGCAACTACGAGAAGATCAACTACAAGCTGCTTCCCGGCCCCAAC
AAGATGCTGCCCAAGGTGTTCTTCTCTAAGAAGTGGATGGCCTACTACAACCCCTCTGA
GGACATCCAGAAGATCTACAAGAACGGCACCTTCAAGAAGGGCGACATGTTCAACCTGA
ACGACTGCCACAAGCTGATCGACTTCTTCAAGGACTCTATCTCTCGATACCCCAAGTGG
TCTAACGCCTACGACTTCAACTTCTCTGAGACCGAGAAGTACAAGGACATCGCCGGCTT
CTACCGAGAGGTGGAGGAGCAGGGCTACAAGGTGTCTTTCGAGTCTGCCTCTAAGAAGG
AGGTGGATAAGCTGGTGGAGGAGGGCAAGCTCTACATGTTCCAGATCTACAACAAGGAC
TTCTCTGACAAGTCTCACGGCACCCCCAACCTGCACACCATGTACTTCAAGCTCCTGTT
CGACGAGAACAACCACGGCCAGATCCGACTGTCTGGCGGCGCCGAGCTGTTCATGCGAC
GAGCCTCTCTGAAGAAGGAGGAGCTGGTGGTGCACCCCGCCAACTCTCCCATCGCCAAC

AAGAACCCCGACAACCCCAAGAAGACCACCACCCTGTCTTACGACGTGTACAAGGACAA
GCGATTCTCTGAGGACCAGTACGAGCTGCACATCCCCATCGCCATCAACAAGTGCCCCA
AGAACATCTTCAAGATCAACACCGAGGTGCGAGTGCTGCTGAAGCACGACGACAACCCC
TACGTGATCGGCATCGACCGAGGCGAGCGAAACCTGCTGTACATCGTGGTGGTGGACGG
CAAGGGCAACATCGTGGAGCAGTACTCTCTGAACGAGATCATCAACAACTTCAACGGCA
TCCGAATCAAGACCGACTACCACTCTCTGCTGGACAAGAAGGAGAAGGAGCGATTCGAG
GCCCGACAGAACTGGACCTCTATCGAGAACATCAAGGAGCTGAAGGCCGGCTACATCTC
TCAGGTGGTGCACAAGATCTGCGAGCTGGTGGAGAAGTACGACGCCGTGATCGCCCTGG
AGGACCTGAACTCTGGCTTCAAGAACTCTCGAGTGAAGGTGGAGAAGCAGGTGTACCAG
AAGTTCGAGAAGATGCTGATCGACAAGCTGAACTACATGGTGGACAAGAAGTCTAACCC
CTGCGCCACCGGCGGCGCCCTGAAGGGCTACCAGATCACCAACAAGTTCGAGTCTTTCA
AGTCTATGTCTACCCAGAACGGCTTCATCTTCTACATCCCCGCCTGGCTGACCTCTAAG
ATCGACCCCTCTACCGGCTTCGTGAACCTGCTGAAGACCAAGTACACCTCTATCGCCGA
CTCTAAGAAGTTCATCTCTTCTTTCGACCGAATCATGTACGTGCCCGAGGAGGATCTGT
TCGAGTTTGCCCTGGACTACAAGAACTTCTCCCGAACCGACGCCGACTACATCAAGAAG
TGGAAGCTGTACTCTTACGGCAACCGAATCCGAATCTTCCGAAACCCCAAGAAGAACAA
CGTGTTCGACTGGGAGGAGGTGTGCCTGACCTCTGCCTACAAGGAGCTGTTCAACAAGT
ACGGCATCAACTACCAGCAGGGCGACATCCGAGCCCTGCTGTGCGAGCAGTCTGACAAG
GCCTTCTACTCTTCTTTCATGGCCCTGATGTCTCTGATGCTGCAGATGCGAAACTCTAT
CACCGGCCGAACCGACGTGGACTTCCTGATCTCTCCCGTGAAGAACTCTGACGGCATCT
TCTACGACTCTCGAAACTACGAGGCCCAGGAGAACGCCATCCTGCCCAAGAACGCCGAC
GCCAACGGCGCCTACAACATCGCCCGAAAGGTGCTGTGGGCCATCGGCCAGTTCAAGAA
GGCCGAGGACGAGAAGCTGGACAAGGTGAAGATCGCCATCTCTAACAAGGAGTGGCTGG
AGTACGCCCAGACCTCTGTGAAGCAC

2.8.2.2 Nucleotide sequence of the sgRNA expression cassette for LbCpf1

To clone in a new gRNA, pCpf1_yl is digested with SpeI and a new 23-25 nt sgRNA is inserted with a ~60 nt primer using Gibson Assembly. On the nucleotide sequence for the gRNA expression cassette, the bolded sequence corresponds to the SCR1'-tRNA^{gly} hybrid PolIII promoter, the italicized sequence corresponds to the direct repeat (DR), the underlined portion references the guide RNA to be cloned into the SpeI digested site, and the sequence that is both bolded and italicized is the PolyT terminator.

CCCCAGTTGCAAAAGTTGACACAACTCTAGATCTGCTTCCAAATATAGAATCATAACAA
GGGTTAGGGTGTGATTATATAATATTGGTCTTAATTGATGTGCTAGGGCTTTAAAAGTT
GGTTAAAATAACGCTCTAATGCCTTTTTAATATATTGTCTTTTTCAAAATCTCAAATCG

GACACTTCTTCGTGTATGAGACTCCATTTTTTGGCTCCGTCACGTGATATGTATTATCA
GCTATAGTGGTGTAAACAAAGTTTTTTACTAGCTGTAATGGCATTTTGTCGGAGTGGTA
AATCGCCTTCTTGTTGTGCGTTCGAGTTCTGGACTCTGCACTGGGCTACTTTGAAAAAT
ACCTCTAATGCGCCGATGGTTTAGTGGTAAAATCCATCGTTGCCATCGATGGGCCCCCG
GTTCGATTCCGGGTCGGCGCA*AATTTCTACTAAGTGTAGAT*NNNNNNNNNNNNNNNNNNN
NNNNNNN***TTTTTTTT***

2.8.3 Supplementary Figures

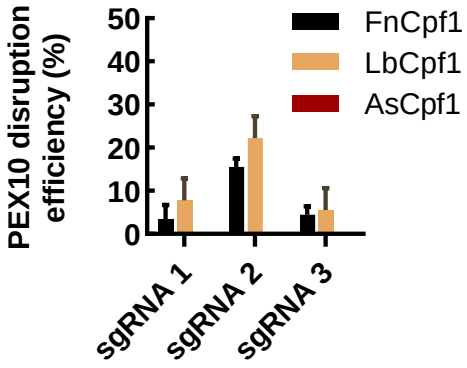

Figure S2.1. Screening activity of Cas12a endonucleases in *Y. lipolytica*. Fn- Lb- and AsCpf1 were tested for nuclease activity in *Y. lipolytica* by disrupting the PEX10 gene. PEX10 is involved in peroxisome biogenesis and is required for *Y. lipolytica* to metabolize long chain fatty acids. PEX10 disruptants were identified by growth on YPD but not on oleic acid. With the same PAM sequence, the same 20 nt gRNA were designed and tested with each Cpf1 variant. Out of the gRNA tested, LbCpf1 showed the best cutting activity (22 ± 5%, n=3), and was chosen for further testing. Note, AsCpf1 did not produce positive results with any of the three gRNAs

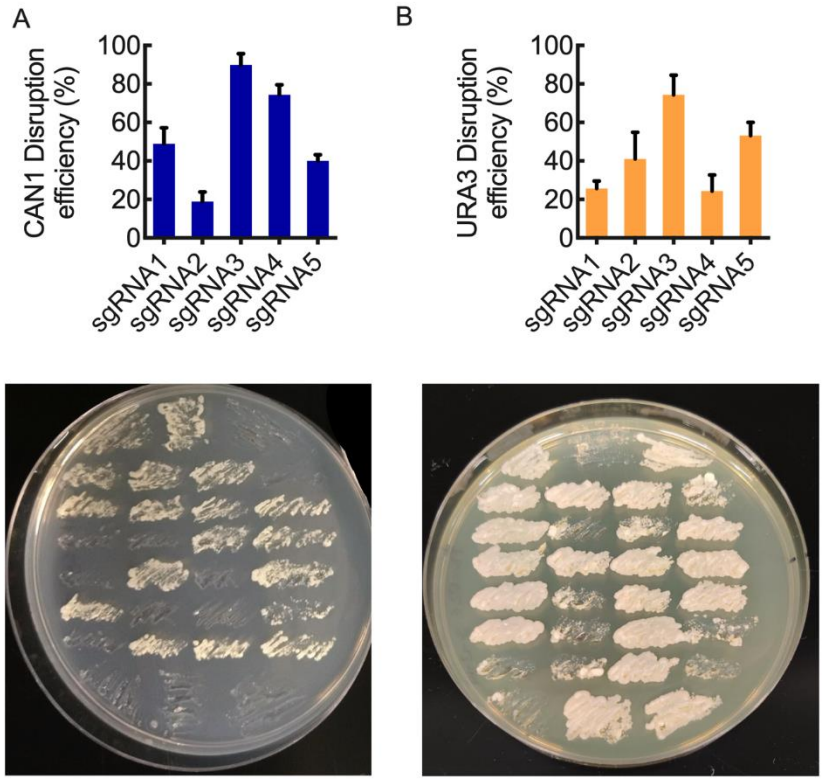

Figure S2.2. Dependence of disruption efficiencies on gRNA for CAN1. (A) and URA3 (B). Examples of plate screening assays are shown below the quantitative data. CAN1 knockout enables growth in plates supplemented with canavanine. URA3 knockouts are able to grow on plates containing 5-FOA.

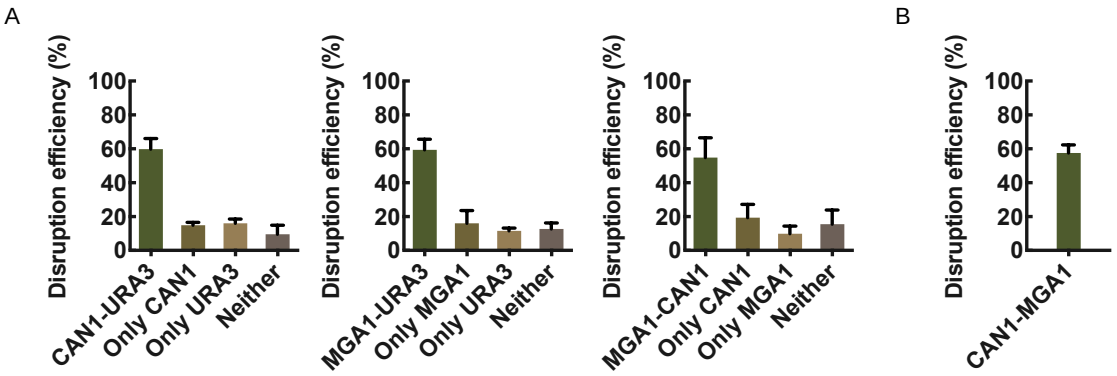

Figure S2.3. Multiplex genome editing using LbCpf1. (A) Dual knockout efficiency with LbCpf1 targeting CAN1/URA3, MGA1/URA3, and CAN1/MGA1. Triplicate experiments (60 colonies per experiment) were screened for dual knockouts. (B) Reversing the configuration of multiplexed CAN1-MGA1 gRNA shows no significant difference in generation of dual knockouts.

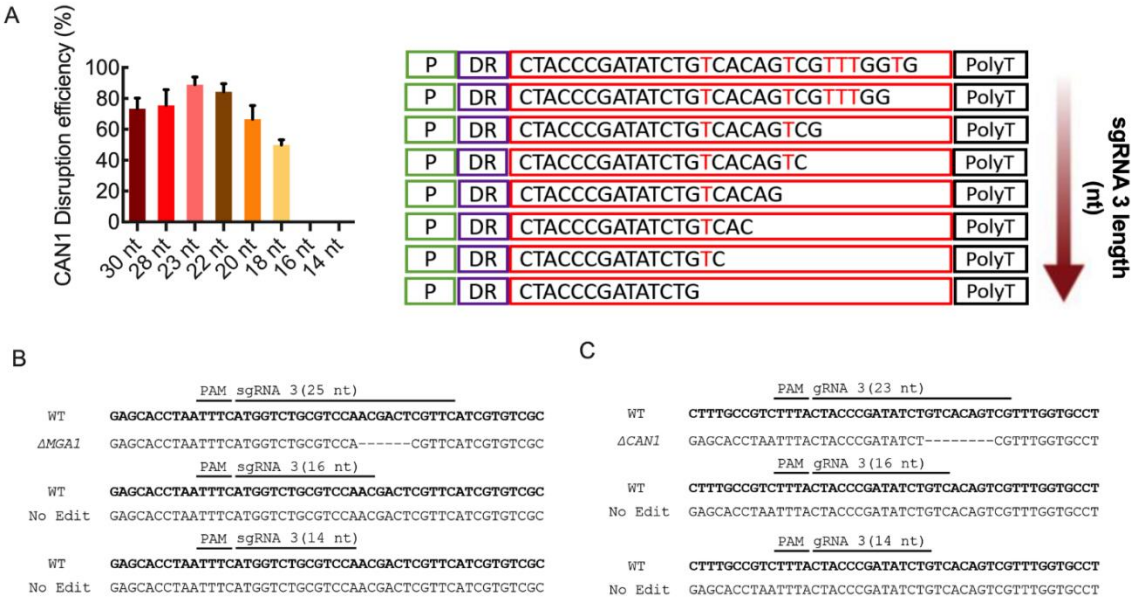

Figure S2.4. **Effect of gRNA length on gene disruption efficiency.** (A) Different lengths of the best gRNA for CAN1 (see Figure S2) from 30 to 14 nt were evaluated for gene disruption. Similar to the MGA1 study (Figure 1D), cutting function is lost with guides of length less than 16 nt. (B,C) Representative examples of MGA1 and CAN1 gene sequence after targeting with truncated, non-functional gRNAs. Thymine (T) nucleotides shown in red indicate locations within each spacer where truncations are not made due to the presence of the polyT terminator.

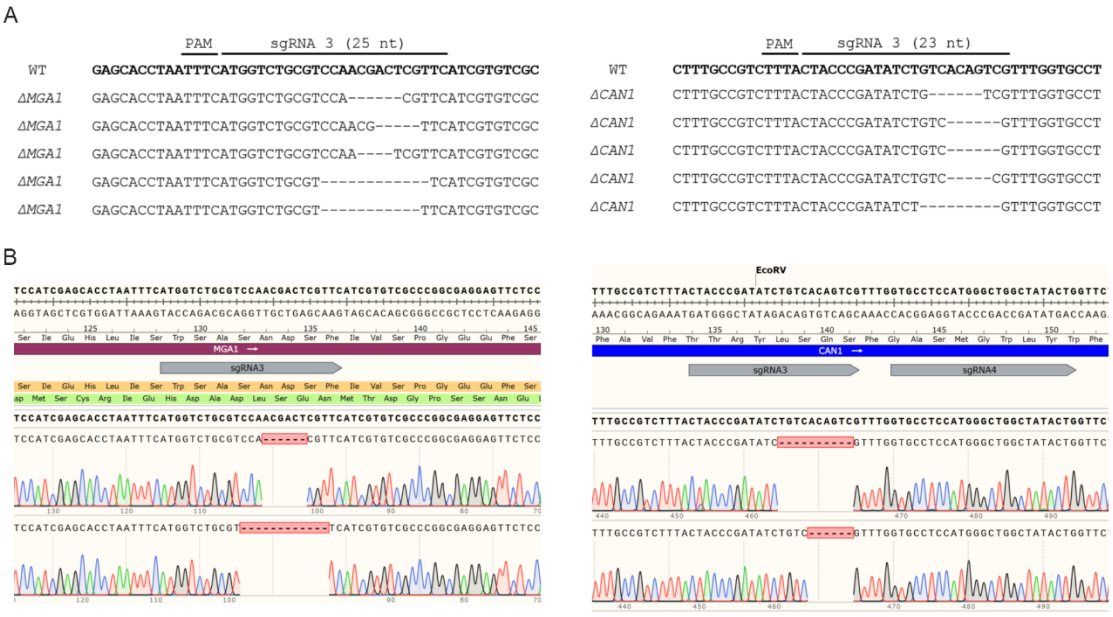

Figure S2.5. Representative sequencing results of MGA1 and CAN1 genes targeted with full length gRNAs. (A) Five colonies showing the phenotype for the null mutant of each gene were sent for sequencing. Deletions are observed around the 18-23 bp as is characteristic of Cpf1 endonuclease activity followed by NHEJ repair. (B) Representative sequence traces and chromatograms showing indels for MGA1 and CAN1.

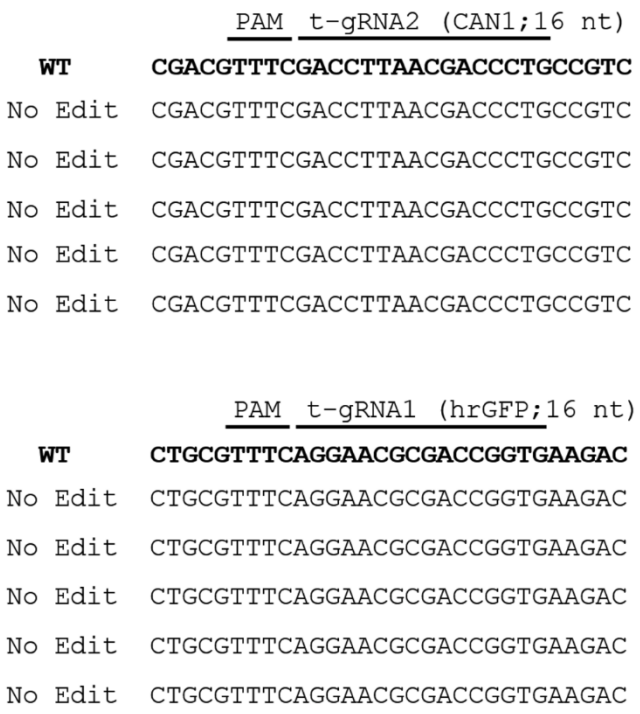

Figure S2.6. Representative sequencing results of CAN1 and hrGFP promoters targeted with truncated gRNAs. For CAN1, t-gRNA2 showed the best CRISPRi repression. Five colonies were sent for sequencing to confirm that there was no Cpf1 nuclease activity and repression was not caused by modifications to the promoter sequence. Similarly, t-gRNA1 showed the most upregulation of hrGFP using CRISPRa. 5 colonies sent for sequencing did not show any edits.

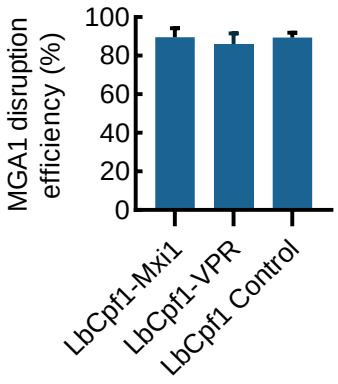

Figure S2.7. Effect of transcriptional regulator fusions to LbCpf1 on gene disruption efficiency. LbCpf1 fused with either a C-terminal Mxi1 repression domain or a C-terminal VPR activation was targeted to MGA1. MGA1 disruption efficiencies achieved were similar in each case. It was thus shown that the C-terminal fusions did not impact nuclease activity of LbCpf1 in any significant manner.

2.8.4 Supplementary Tables

Table S2.1. Method comparison to contemporary CRISPR-Cpf1 tools in *Yarrowia lipolytica*

	Zhang et al.[10]	Yang et al.[11]	This study
Cpf1 endonuclease used	FnCpf1	AsCpf1	LbCpf1
Direct repeat	19	20	20
Genome editing	No	Yes	Yes
Singleplex efficiency	-	40-96%	75-90%
Multiplexing	-	3	3
Multiplex efficiency	-	42-83%	44-60%
Gene Regulation	CRISPRi	N/A	CRISPRi/a
Multiplexing	3	-	2
Technology used	dFnCpf1	-	LbCpf1/ truncated gRNA
Simultaneous gene disruption and regulation	-	-	Yes

Table S2.2. Yeast strains used in this study.

Yeast strain genotype	Phenotype
PO1f	Wild type strain
PO1f URA3::A08	PO1f expressing URA3 gene at A08 locus, alleviating uracil auxotrophy.
PO1f GAL1 UAS-TEF-hrGFP::XPR2	PO1f expressing hrGFP from a TEF core promoter and GAL1 upstream activation sequence, for minimal GFP expression.

Table S2.3. Sequences of primers used in this study

Primer name	Primer Sequence
SCR_DR F	CCGAAAAGTGCCACCTGACGTCCCCAGTTGCAAAAGTTGACAC
SCR_DR_R_Lb	GATAATAATGGTTTCTTAGACGTAAAAAACTAGTCTACACTTAGTAGAAATTTGCGCCGACCCCGGAATC
SCR_DR-R_Fn	GATAATAATGGTTTCTTAGACGTAAAAAACTAGTCTACAACAGTAGAAATTTGCGCCGACCCGGAATCGAAC
PEX_Sg1_Fn	AATTTCTACTGTTGTAGATGATTGTCGTATTGTCGCTCATTTTTTTTACGTCTAAGAAAC
PEX_Sg2_Fn	AATTTCTACTGTTGTAGATTCCACCAGTACAAGGAGGAGTTTTTTACGTCTAAGAAAC
PEX_Sg3_Fn	AATTTCTACTGTTGTAGATCATATCTCGGTTTGTGTACGTTTTTTTTACGTCTAAGAAAC
PEX_Sg1_Lb	ATTTCTACTAAGTGTAGATGATTGTCGTATTGTCGCTCATTTTTTTTACGTCTAAGAAAC
PEX_Sg2_Lb	ATTTCTACTAAGTGTAGATTCCACCAGTACAAGGAGGAGTTTTTTACGTCTAAGAAAC
PEX_Sg3_Lb	ATTTCTACTAAGTGTAGATCATATCTCGGTTTGTGTACGTTTTTTTTACGTCTAAGAAAC
MGA_Sg1	ATTTCTACTAAGTGTAGATGGCGGCATGTGCTCGACCCGTTCTTTTTTACGTCTAAGAAA
MGA_Sg2	TTTCTACTAAGTGTAGATGAGTGGTGCCGGGCTTCTTGTTATCTTTTTTACGTCTAAGAA
MGA_Sg3	TTTCTACTAAGTGTAGATATGGTCTGCGTCCAACGACTCGTTCTTTTTTACGTCTAAGAA
MGA_Sg4	ATTTCTACTAAGTGTAGATTGCGCCAGCTCAACATGTACGGCTTTTTTACGTCTAAGAAA
MGA_Sg5	ATTTCTACTAAGTGTAGATCACACCGGCGACTCCTCGCAATGGTTTTTTACGTCTAAGAA
CAN_Sg1	ATTTCTACTAAGTGTAGATAAACGATTACCCACCCTCCGGGACTTTTTTACGTCTAAGAA
CAN_Sg2	TTTCTACTAAGTGTAGATCTTGTGCGAGGGCACCTCCTCTGAGTTTTTTACGTCTAAGAA
CAN_Sg3	ATTTCTACTAAGTGTAGATCTACCCGATATCTGTCACAGTCGTTTTTTACGTCTAAGAAA
CAN_Sg4	TTTCTACTAAGTGTAGATGTGCCTCCATGGGCTGGCTATACTGTTTTTTACGTCTAAGAA
CAN_Sg5	TTTCTACTAAGTGTAGATGCACAATGGGCACGCCGTCGGTCCATTTTTTACGTCTAAGAA
URA_Sg1	TTTCTACTAAGTGTAGATCCGCTCGAGTGCTCAAGCTCGTGGCTTTTTTACGTCTAAGAA
URA_Sg2	TTTCTACTAAGTGTAGATTGTCCTCGAACAGGAAGAAACCGTGTTTTTTTACGTCTAAGAA
URA_Sg3	TTTCTACTAAGTGTAGATCTCGGCACCAGCTCGCAGGCCAGCATTTTTTTACGTCTAAGAA
URA_Sg4	TTTCTACTAAGTGTAGATTTCCTGTTCGAGGACAGAAAGTTCGTTTTTTTACGTCTAAGAA
URA_Sg5	TTTCTACTAAGTGTAGATTTGGCTGCCACGAGCTTGAGCACTCTTTTTTTACGTCTAAGAA
MGA_MC	GGCGCATAATTTCTACTAAGTGTAGATATGGTCTGCGTCCAACGACTCGTTCAATTTCTACTAAGTGTAGATCTACCCGATATCT
CAN_MC	ATGGTTTCTTAGACGTAAAAAACGACTGTGACAGATATCGGGTAGATCTACACTTAGTAGAAATTGAACGAGTCGTTGGA
MGA_MU	GGCGCATAATTTCTACTAAGTGTAGATATGGTCTGCGTCCAACGACTCGTTCAATTTCTACTAAGTGTAGATCTCGGCACCAGCT
URA_MU	TTCTTAGACGTAAAAAATGCTGGCCTGCGAGCTGGTGCCGAGATCTACACTTAGTAGAAATTGAACGAGTCGTTGGA
CAN_CU	GGCGCATAATTTCTACTAAGTGTAGATCTACCCGATATCTGTCACAGTCGAATTTCTACTAAGTGTAGATCTCGGCACCAGCT
URA_CU	TTCTTAGACGTAAAAAATGCTGGCCTGCGAGCTGGTGCCGAGATCTACACTTAGTAGAAATTCGACTGTGACAGATA
CMU_1	CAAATTTCTACTAAGTGTAGATCTACCCGATATCTGTCACAGTCGAATTTCTACTAAGTGTAGATATGGTCTGCGTCCAACGACTCGTTC
CMU_2	GGTTTCTTAGACGTAAAAAATGCTGGCCTGCGAGCTGGTGCCGAGATCTACACTTAGTAGAAATTGAACGAGTCGTTGGACGCAGACCAT
MGA_31	TTTCTACTAAGTGTAGATATGGTCTGCGTCCAACGACTCGTTCATCGTGTTTTTTACGTCTAAGAA
MGA_28	TTCTACTAAGTGTAGATATGGTCTGCGTCCAACGACTCGTTCATCTTTTTTACGTCTAAG
MGA_22	TTTCTACTAAGTGTAGATATGGTCTGCGTCCAACGACTCGTTTTTTACGTCTAAGAA
MGA_19	TTTCTACTAAGTGTAGATATGGTCTGCGTCCAACGACTTTTTTTACGTCTAAGAA

MGA_18	TTTCTACTAAGTGTAGATATGGTCTGCGTCCAACGATTTTTTACGTCTAAGAA
MGA_16	TTTCTACTAAGTGTAGATATGGTCTGCGTCCAACTTTTTTACGTCTAAGAA
MGA_14	TTTCTACTAAGTGTAGATATGGTCTGCGTCCATTTTTTACGTCTAAGAA
CAN_30	ATTTCTACTAAGTGTAGATCTACCCGATATCTGTCACAGTCGTTTGGTGTTTTTTACGTCTAAGAAA
CAN_28	ATTTCTACTAAGTGTAGATCTACCCGATATCTGTCACAGTCGTTTGGTTTTTTACGTCTAAGAAA
CAN_22	ATTTCTACTAAGTGTAGATCTACCCGATATCTGTCACAGTCTTTTTTACGTCTAAGAAA
CAN_20	ATTTCTACTAAGTGTAGATCTACCCGATATCTGTCACAGTTTTTTACGTCTAAGAAA
CAN_18	ATTTCTACTAAGTGTAGATCTACCCGATATCTGTCACTTTTTTACGTCTAAGAAA
CAN_16	ATTTCTACTAAGTGTAGATCTACCCGATATCTGTCTTTTTTACGTCTAAGAAA
CAN_14	ATTTCTACTAAGTGTAGATCTACCCGATATCTGTTTTTTACGTCTAAGAAA
CAN_Tsg1	ATTTCTACTAAGTGTAGATCATTGTGGTCGGATGGTTTTTTACGTCTAAGAAA
CAN_Tsg2	ATTTCTACTAAGTGTAGATGACCTTAACGACCCTGTTTTTTACGTCTAAGAAA
CAN_Tsg3	ATTTCTACTAAGTGTAGATGTGGGGAGCGTCGTCCTTTTTTACGTCTAAGAAA
CAN_Tsg4	ATTTCTACTAAGTGTAGATATGGAATCTGATGTGGTTTTTTACGTCTAAGAAA
CAN_Tsg5	ATTTCTACTAAGTGTAGATTGCCCTTCAAAACCAGTTTTTTACGTCTAAGAAA
GFP_Tsg1	ATTTCTACTAAGTGTAGATAGGAACGCGACCGGTGTTTTTTACGTCTAAGAAA
GFP_Tsg2	ATTTCTACTAAGTGTAGATTACAATTGCGGAGCAGTTTTTTACGTCTAAGAAA
GFP_Tsg3	ATTTCTACTAAGTGTAGATTCTCTCTCCTTGTCAATTTTTTACGTCTAAGAAA
GFP_Tsg4	ATTTCTACTAAGTGTAGATGGGTGTGAGTTGACAATTTTTTACGTCTAAGAAA
GFP_Tsg5	ATTTCTACTAAGTGTAGATCTTCTGAGTATAAGAATTTTTTACGTCTAAGAAA
GFP_Tsg6	ATTTCTACTAAGTGTAGATAATGATTCTTATACTCTTTTTTACGTCTAAGAAA
Mxi-F	GACCCCAAGAAGAAGCGAAAGGTGGGTGGATCTGGTGGATCTGGCTCCTCTAAGCTGGGC
Mxi-R	AGGGCGTGAATGTAAGCGTGAC
VPR-F	TCTCGAGCCGACCCCAAG
VPR-R	TTCGGTTAGAGCGGATGTGG
qCAN-F	CATTGGTCCCGTGATTGAG
qCAN-R	GGGAAGAAGTTGATGGTAGTG
qGFP-F	AACAGCGGCAAGTTCTAC
qGFP-R	CGGTGCTGGATGAAGTG
Act-F	TCCAGGCCGTCCTCTCCC
Act-R	GGCCAGCCATATCGAGTCGCA
Sg-Seq	CTTCGACTCTAGAGGATCTGG
dCpf1_SDM_F	ATCGGCATCGCCCGAGGCGAG
dCpf1_SDM_R	CACGTAGGGGTTGTCGTCG
Non-Targeting	ATTTCTACTAAGTGTAGATCCGCTGTGTAGCGGACTTTTTTACGTCTAAGAAA
CAN-Tsg2-full	TTTCTACTAAGTGTAGATGACCTTAACGACCCTGCCGTCTCCATTTTTTACGTCTAAGAA
GFP-Tsg1-full	TTTCTACTAAGTGTAGATAGGAACGCGACCGGTGAAGACGAGGTTTTTTACGTCTAAGAA

2.8.5 References

1. Madzak, C.; Tréton, B.; Blanchin-Roland, S., Strong hybrid promoters and integrative expression/secretion vectors for quasi-constitutive expression of heterologous proteins in the yeast Yarrowia lipolytica. *J. Mol. Microbiol. Biotechnol.* **2000,** *2* (2), 207-216.

2. Schwartz, C.; Curtis, N.; Lobs, A. K.; Wheeldon, I., Multiplexed CRISPR Activation of Cryptic Sugar Metabolism Enables Yarrowia Lipolytica Growth on Cellobiose. *Biotechnol. J.* **2018,** *13* (9).

3. Schwartz, C.; Frogue, K.; Ramesh, A.; Misa, J.; Wheeldon, I., CRISPRi repression of nonhomologous end-joining for enhanced genome engineering via homologous recombination in Yarrowia lipolytica. *Biotechnol. Bioeng.* **2017,** *114* (12), 2896-2906.

4. Kleinstiver, B. P.; Tsai, S. Q.; Prew, M. S.; Nguyen, N. T.; Welch, M. M.; Lopez, J. M.; McCaw, Z. R.; Aryee, M. J.; Joung, J. K., Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nat. Biotechnol.* **2016,** *34* (8), 869.

5. Puigbo, P.; Guzman, E.; Romeu, A.; Garcia-Vallve, S., OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res.* **2007,** *35* (suppl_2), W126-W131.

6. Zetsche, B.; Gootenberg, J. S.; Abudayyeh, O. O.; Slaymaker, I. M.; Makarova, K. S.; Essletzbichler, P.; Volz, S. E.; Joung, J.; Van Der Oost, J.; Regev, A., Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **2015,** *163* (3), 759-771.

7. Gibson, D. G.; Young, L.; Chuang, R.-Y.; Venter, J. C.; Hutchison III, C. A.; Smith, H. O., Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **2009,** *6* (5), 343.

8. Schwartz, C. M.; Hussain, M. S.; Blenner, M.; Wheeldon, I., Synthetic RNA Polymerase III Promoters Facilitate High-Efficiency CRISPR-Cas9-Mediated Genome Editing in Yarrowia lipolytica. *ACS Synth. Biol.* **2016,** *5* (4), 356-359.

9. McMillan, J.; Lu, Z.; Rodriguez, J. S.; Ahn, T.-H.; Lin, Z., YeasTSS: an integrative web database of yeast transcription start sites. *Database* **2019,** *2019*.

10. Zhang, J.-l.; Peng, Y.-Z.; Liu, D.; Liu, H.; Cao, Y.-X.; Li, B.-Z.; Li, C.; Yuan, Y.-J., Gene repression via multiplex gRNA strategy in Y. lipolytica. *Microb. Cell Fact.* **2018,** *17* (1), 62.

11. Yang, Z.; Edwards, H.; Xu, P., CRISPR-Cas12a/Cpf1-assisted precise, efficient and multiplexed genome-editing in Yarrowia lipolytica. *Metab. Eng. Commun.* **2020,** *10*, e00112.

Chapter 3: Guide RNA design for genome-wide CRISPR Screens in *Yarrowia lipolytica*

3.1 Abstract

Genome-wide functional genomic screens are essential to determining the genetic underpinning of a biological process. Novel and powerful tools for perturbing gene function, with the help of genetic and epigenetic information have made it possible to systematically investigate the contribution of every gene to evolved and engineered phenotypes. Functional genomics and screening for enhanced phenotypes become ever more important when dealing with non-conventional hosts. Non-model organisms are valuable to metabolic engineering as they present a range of desirable phenotypes and can help in avoiding complex and intensive engineering of less suitable hosts that do not possess the desired phenotype(s). Domestication of such hosts however requires a suite of synthetic biology tools that allow for targeted genome engineering, regulation of gene expression, and critically genome-wide mutational screens. The widespread adoption of CRISPR-Cas9 and CRISPR-Cpf1 based systems has allowed for such screens in many organisms. Key considerations in any genome-wide CRISPR screen are the design of a set of unique guide-RNA targeting the required set of genes in the genome, and the design of non-targeting guide-RNA that function as appropriate negative controls for the experiment. In this methods chapter, we present a protocol for the design of guides for a CRISPR screen, targeting every gene in the genome of the industrially relevant oleaginous yeast *Yarrowia lipolytica*. The first set of protocols describe the algorithm for the design of genome targeting and non-targeting guides for a genome-wide CRISPR-Cpf1 screen. The second

set of protocols describes modifications to the first for the design of guides for a CRISPR-Cas9 screen. The strategies described here should serve as an efficient guide to design a library of gRNA for most genome-wide CRISPR screens.

This chapter previously appeared as a method chapter for *Yarrowia lipolytica* in the *Methods in Molecular Biology* book series Volume 2307. The original citation is as follows: Ramesh, A., & Wheeldon, I. (2021). Guide RNA design for genome-wide CRISPR screens in Yarrowia lipolytica. In *Yarrowia lipolytica* (pp. 123-137). Humana, New York, NY.

3.2 Introduction

The goal of functional genomics is to use the diverse information obtained from genomes, to perturb gene function and determine the genetic underpinnings of the resulting phenotype. In the postgenomic era, with advances in DNA sequencing and synthesis, and synthetic biology we have the genetic information and capacity to perform genome-wide screens by systematic loss-of-function studies. Pooled forward genetic screens are puissant tools that help discover genes that affect a desired phenotype, by facilitating a different genetic perturbation in each cell prior to a selection based on a required phenotype [1]. The cornerstone of pooled genetic screens is that the selection pressure applied to select for a desirable phenotype, results in a distribution in the occurrence of genetic perturbation events that affect that phenotype. The genetic perturbations that were most enriched or depleted may then be investigated as promising gene targets to enhance the phenotype. With the advent of CRISPR technologies for genome engineering, new opportunities for performing pooled genetic screens by modifying DNA in a targeted manner have arisen [2-4]. It is now possible to track the generated gene edits, with the added advantage of being able to perform targeted site saturation mutagenesis [5,6]. While broadly applied to various genomic studies, such genome-wide screening techniques have yet to be extended to many non-model and other industrially relevant hosts, that would greatly benefit from such a screen.

Non-model organisms often make valuable hosts for bioprocessing due to their natural capability to produce a desired product, as well as their possession of beneficial native phenotypes [7]. While this helps avoid complex and intensive engineering of unsuitable

hosts, non-model hosts may carry the challenge of a lack of synthetic biology tools and complete biological understanding [8]. Domestication of such hosts requires novel synthetic biology tools for generation of targeted gene knock-outs, integration of heterologous pathways, regulation of gene transcription and most importantly, genome-scale mutational screens to enhance the desired phenotype and investigate the underlying genetic elements.

The non-model yeast *Yarrowia lipolytica* has garnered attention as a eukaryotic host for metabolic engineering due to its capability to grow on diverse hydrocarbon substrates and accumulate high levels of intracellular lipids [9-11]. This oleaginous yeast also finds industrial relevance as a bioprocessing host for conversion of biomass derived sugars and industrial waste into lipid, and lipid-derived, as well as other value-added products [12-15]. The easy programmability of CRISPR-Cas9 based systems has expedited genetic engineering in *Y. lipolytica*. Novel CRISPR tools capable of targeted gene knockouts, standardized sites for markerless gene integration, and regulation of gene expression with CRISPRi and CRISPRa systems have furthered the prospects of convenient strain engineering with this yeast [16-19]. However there exist bottle necks in rapid strain development using *Y. lipolytica*, such as complications with multi-gene editing efficiencies, precise site-directed mutagenesis, and until recently genome-scale mutational screens for functional genomics [20]. Cas12a endonucleases, also known as Cpf1 endonucleases, present themselves as attractive alternatives or orthogonal complements to gene editing with CRISPR-Cas9 [21]. Cpf1 doesn't require a tracrRNA for gene editing and harbors an endoribonuclease domain capable of maturing its own CRISPR-RNA array

into individual sgRNA. These qualities drive Cpf1 as an appealing system for multiplex gene editing.

Schwartz et al., in 2019 described a strategy for quantitatively validating the genome wide function of a synthesized sgRNA library, by designating a cutting efficiency score every sgRNA in the library [20]. This helped distinguish active from inactive guides in the library, and to quantify genome wide coverage, by revealing the presence of false negatives among sgRNA. While prediction software for sgRNA activity are emerging [22,23], there is still insufficient data that can correlate Cas enzyme activity, sgRNA sequence, local genetic and epigenetic features to provide a consistent tool for sgRNA design.

In this protocol chapter, we provide a method to design a sgRNA library with n-fold coverage of all genes in *Y. lipolytica*, for performing genome-scale CRISPR-Cas9 and CRISPR-Cpf1 screens. This library can then be validated as described in Schwartz et al. (2019) [20] in order to obtain a list of known active guides for each gene that may be taken to following screens.

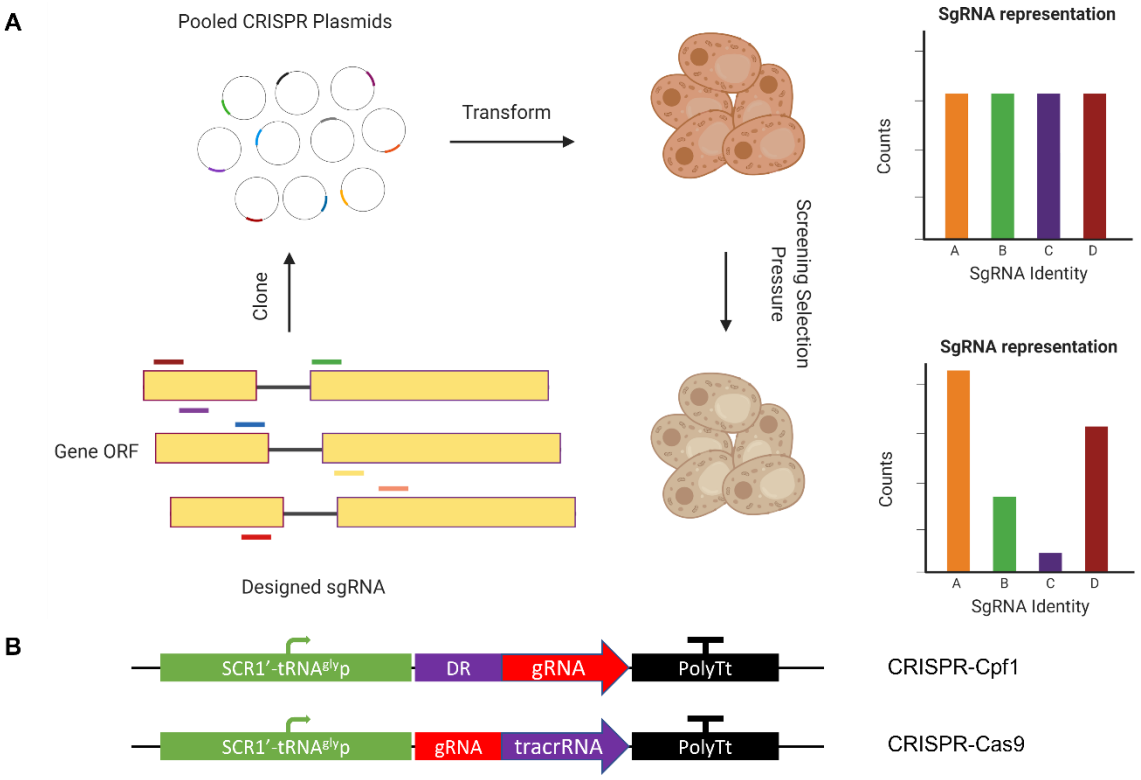

Figure 3.1. Schematic of pooled CRISPR genome-wide screens. (A) Guide RNA are designed for every ORF in the genome, cloned, and transformed into the strain of interest. The selection pressure applied causes a perturbation in sgRNA abundance which can be measured to draw conclusions on the importance of the gene it targets. (B) Schematic of the sgRNA expression cassette for CRISPR-Cpf1 (top) and CRISPR-Cas9 (bottom) gene editing. sgRNAs are expressed from a hybrid PolIII-tRNA promoter with a PolyT terminator. CRISPR-Cpf1 gene editing requires a 20 nt Direct Repeat (DR) followed by a 23-25 nt gRNA sequence, while CRISPR-Cas9 systems require a 20 nt gRNA followed by a 89 nt tracrRNA sequence.

3.3 Materials

3.3.1 Software and computer

1. Mathworks MATLAB version R2018b or later

2. MATLAB bioinformatics toolbox

3. Laptop or desktop computer that meets the requirements to run MATLAB R2018b

3.4 Methods

3.4.1 CRISPR Plasmids for CRISPR-Cas9 and CRISPR-Cpf1 Gene Editing

1. In our CRISPR related gene editing protocols, the CRISPR plasmids express the Cas9 or Cpf1 endonuclease from a high expression UAS1B8-TEF promoter [24,25]. The endonuclease is also fused with a C-terminal SV40 nuclear localization sequence, for import of the protein into the cell nucleus. The plasmids also contain the ampicillin resistance marker (AmpR) for maintenance in *E. coli*, and the LEU2 marker for maintenance and selection in *Yarrowia lipolytica*.

2. The CRISPR plasmids also contain the expression cassettes for the single guide RNA (sgRNA), that targets the endonuclease to genome for gene editing. The sgRNA is expressed from a hybrid PolIII promoter, wh ich combines the native class II Pol III promoter SCR1, with tRNA for glycine. This allows for the maturation and excision of the sgRNA from the primary transcript.

3. While the general structure of the CRISPR plasmid remains the same for gene editing with Cas9 and Cpf1, the sgRNA expression cassette itself has slight variations. In the

case of Cas9, the SCR1'-$tRNA^{gly}$ Pol III promoter is followed by the 20 bp genomic target sequence (spacer), and an 79 bp tracrRNA sequence which functions as a handle for the Cas9. Transcription is terminated with the help of a polyT terminator (TTTTTT) which makes for 99 bp sgRNA transcript.

4. In the case of the Cpf1 CRISPR plasmid, the sgRNA is expressed from the SCR1'-$tRNA^{gly}$ Pol III promoter much like its Cas9 counterpart [19]. The sgRNA sequence itself consists of a 20 bp direct repeat sequence, followed by a 23-25 bp spacer sequence. The direct repeat forms a stem loop structure when transcribed and is essential for Cpf1-mediated cleavage of target DNA. No additional tracrRNA is required. Transcription of the sgRNA cassette is again terminated with the help of a polyT terminator, which makes for a 43-45 bp sgRNA transcript.

3.4.2 Design of an sgRNA library for a CRISPR-Cpf1 genome wide screen

The following protocol describes the algorithm for the design of an sgRNA library for a genome-wide CRISPR-Cpf1 screen in Yarrowia lipolytica strain CLIB89(W29). The algorithm for the design of the library spans an n-fold coverage of each gene. The code for the generation of a library with 8-fold coverage was written on the latest MATLAB version with access to the bioinformatics toolbox and is available upon request.

1. On the NCBI website, search for the nucleotide sequences of the yali1 genome. This provides a list of the 6 chromosomes of the CLIB89 strain. Each listed item should have a separate webpage with complete sequence information of the chromosome and its annotated features. From this page, both the full sequence of the chromosome and

the coding features should be downloaded separately and saved as FASTA files. This will leave a set of 6 FASTA files having complete chromosome sequence and 6 FASTA files containing only the coding features of each respective chromosome.

2. On MATLAB, the fastaread command in the bioinformatics toolbox can be taken advantage of to import FASTA sequences as variables. Thus, all coding features from each of the 6 chromosomes can be imported into a single variable and stored with the appropriate headers to identify each gene and its chromosomal location. For the sake of simplicity this variable will be denoted as 'Genes_Topstrand'.

3. The fastaread command is similarly used again to import the complete sequence of each of the 6 chromosomes into another variable, containing the appropriate identifiers. For the sake of simplicity this variable is denoted as 'Chromosomes'.

4. sgRNAs for CRIPSR gene editing may be generated from either strand of the genomic DNA. Since Genes_Topstrand contains only the coding strand information for each gene, we create another variable Genes_Bottomstrand that contains the complementary sequences of each gene in Genes_Topstrand.

5. The next step is to generate all possible sgRNA from the top and bottom strands for each gene in CLIB89. The PAM sequence for gene editing using Cpf1 is 5'-TTTV-3' (V=A/G/C), with the spacer being the following 25 nt. Thus, all TTTV sequences for each gene in Genes_Topstrand and Genes_Bottomstrand are flagged and the following 25 nt are recorded. This may be done in two separate variables Sg_Top and Sg_Bottom.

Both these variables will have as many rows as there are genes (7919) and each row will have varying number of non-empty columns depending on the number of sgRNA identified.

6. Now that we have a list of all possible sgRNA for each gene, we now subject them to a test for uniqueness to sort through and discard any sgRNA that may not be or has a high chance of causing off-target editing. The criteria for uniqueness in Cpf1 sgRNA in our code, was that the first 14 bp each sgRNA be completely unique (Note 1). This means no sgRNA for Cpf1 that cuts anywhere in the entire genome, shares the first 14 bp with another sgRNA.

7. This was achieved by first generating a list of all possible Cpf1 sgRNA in CLIB89. Complementary sequences for each chromosome in the variable Chromsomes were generated and sgRNA were generated from both strands leading to a variable that contained 6 rows and as many non-empty columns for each row as there are sgRNA for that chromosome. To ensure a harsher criterion for uniqueness, the sgRNA generated in this list were preceded by a TTTN (N = A/T/G/C) PAM sequence. These sgRNA were stored in a single variable denoted as Sg_All.

8. The test for uniqueness was conducted by comparing the first 14 bp of every sgRNA in Sg_Top and Sg_Bottom, to the first 14 bp of every single sgRNA in Sg_All. Since Sg_Top and Sg_Bottom are subsets of Sg_All, if any sgRNA in Sg_Top and Sg_Bottom, repeated more than once in the above comparison, that guide was

discarded. This comparison was iterated through all guides in Sg_Top and Sg_Bottom, to generate two new variables, Sg_Top_unique and Sg_Bottom_unique.

9. The final step of this procedure involves picking sgRNA from Sg_Top_unique and Sg_Bottom_unique to create an sgRNA library that has n-fold coverage of each gene. To create a non-biased library, half the generated unique guides may be picked from each strand to make the final library. Our code generates a Cpf1 library with an 8-fold coverage for each gene. If both top and bottom strands for a gene contained more than 4 unique sgRNA, 4 from each strand were picked to make up the final set. If either strand for any gene contained less than 4 unique sgRNA, the remaining would be picked from the other strand. If both strands contained less than 4 unique sgRNA, then all sgRNA would be picked to make the final library.

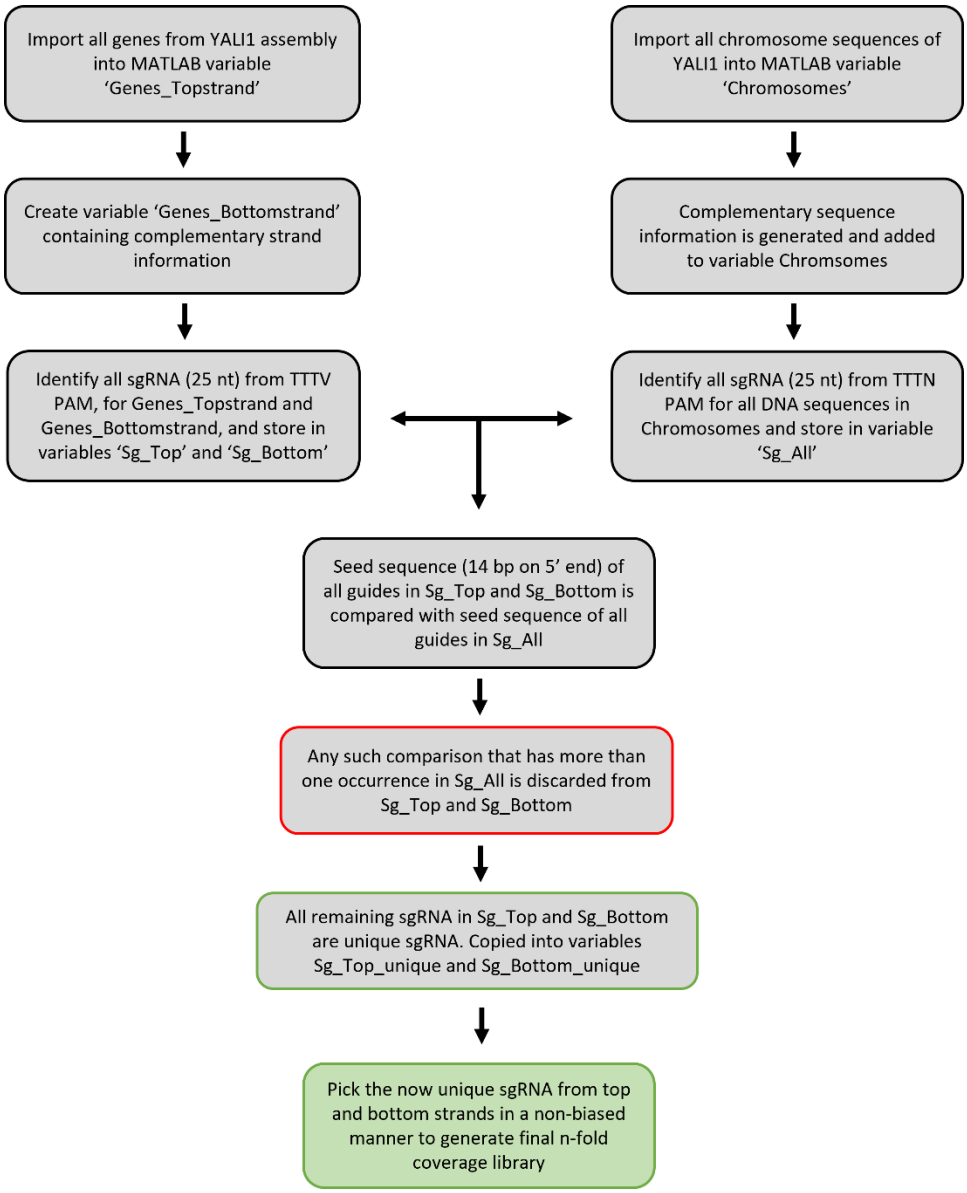

Figure 3.2. Flow diagram for the design of an n-fold coverage library of sgRNA for pooled CRISPR-Cpf1 screens. The library was designed for the CLIB89 (W29) strain of Y. lipolytica. Genes and chromosomes were imported from NCBI as two separate sets, PAM sites identified, and gRNAs flagged for each set before a seed sequence of 14 nt on the 5' end was used to test for uniqueness. Once all non-unique gRNA are eliminated, the remaining can be picked in a non biased manner to make up an n-fold coverage library. An 8-fold coverage library containing 57,771 gRNA was designed for CLIB89 using this method.

3.4.3 Design of an sgRNA library for a CRISPR-Cas9 genome wide screen

The algorithm for the design of an sgRNA library for a CRISPR-Cas9 genome wide screen follows closely with the algorithm discussed above with a few changes. Since there exist scoring algorithms that calculate on target cutting efficiency of Cas9 guides, these may be incorporated into the library design procedure to rank all the generated sgRNA for each gene. This may be of some help in down-selecting and deciding which guides may be picked for each gene.

1. Similar to the previous algorithm, sequence information about the CLIB89 strain is essential. The full sequence of each chromosome and its coding features should be downloaded separately and saved as FASTA files. On MATLAB, the 'fastaread' command can be used to import FASTA sequences as variables. Thus, all coding features from each of the 6 chromosomes can be imported into a single variable and stored with the appropriate headers to identify each gene and its chromosomal location. This variable is denoted as Genes_Topstrand. A variable called Genes_Bottomstrand, containing the complementary strand sequence to all genes in Genes_Topstrand is also created.

2. The fastaread command is similarly used again to import the complete sequence of each of the 6 chromosomes into another variable, along the appropriate identifiers. This variable is denoted as 'Chromosomes'.

3. sgRNAs for CRIPSR gene editing may be generated from either strand of the genomic DNA. Since Genes_Topstrand contains only the coding strand information for each gene, we create another variable Genes_Bottomstrand that contains the complementary sequences of each gene in Genes_Topstrand.

4. The next step is to generate all possible sgRNA from the top and bottom strands for each gene in CLIB89. The PAM sequence for gene editing using Cas9 is 5'-NGG-3' (N=A/T/G/C), with the spacer being the 20 nt upstream of the PAM. NGG is more frequently occurring in the genome than the PAM sequence for Cpf1, which is TTTV. As a result, the Cas9-sgRNA for any gene number far greater than their Cpf1 counterpart. Thus, all NGG sequences for the first 300 bp of each gene in Genes_Topstrand and Genes_Bottomstrand are flagged and the 20 nt immediately upstream are recorded. This may be done in two separate variables Sg_Top and Sg_Bottom. Both these variables will have as many rows as there are genes (7919) and each row will have varying number of non-empty columns depending on the number of sgRNA identified.

5. Complementary sequences for each chromosome in the variable Chromsomes were generated and sgRNA were generated from both strands leading to a variable that contained 6 rows and as many non-empty columns for each row as there are sgRNA for that chromosome. Once again, the sgRNA generated in this list were preceded by a NGG PAM sequence. These sgRNA were stored in a single variable denoted as Sg_All.

6. Similar to the Cpf1 library, we subject the generated sgRNA through a test for uniqueness. This was conducted by comparing the last 12 bp of every sgRNA in Sg_Top and Sg_Bottom, to the last 12 bp of every single sgRNA in Sg_All. Since Sg_Top and Sg_Bottom are subsets of Sg_All, if any sgRNA in Sg_Top and Sg_Bottom, repeated more than once in the above comparison, that guide was discarded. This comparison was iterated through all guides in Sg_Top and Sg_Bottom, to generate two new variables, Sg_Top_unique and Sg_Bottom_unique.

7. At this point, an sgRNA on target score calculator like the one described by Doench et al. (2014) or others [22,26] can be used to score and rank the sgRNA of each gene making it easier to pick targets for a library of n-fold coverage. If the library is made purely based on the rankings of the scoring algorithm, the unique top and bottom strand guides may be combined into a single variable before being subjected to scoring.

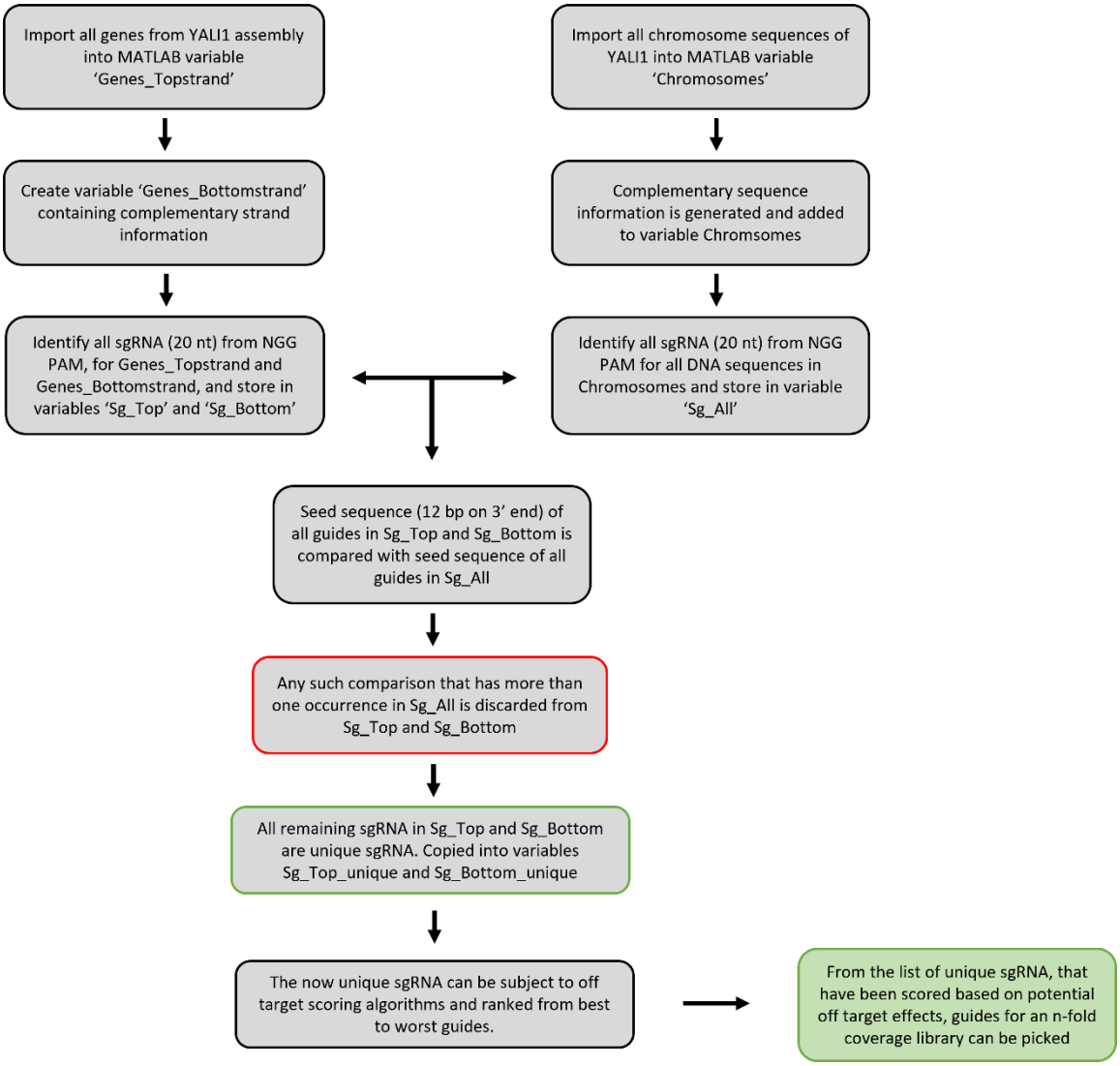

Figure 3.3. Flow diagram for the design of an n-fold coverage library of sgRNA for pooled CRISPR-Cas9 screens. Again, the library was designed for the CLIB89 (W29) strain of Y. lipolytica. Genes and chromosomes were imported from NCBI as two separate sets, PAM sites identified, and gRNAs flagged for each set before a seed sequence of 12 nt on the 3' end was used to test for uniqueness. Once all non-unique gRNA are eliminated, the gRNA can then be subjected to a scoring algorithm to increase probability of picking good gRNA, and the top n/2 gRNA can be picked from each strand to make an unbiased n-fold coverage library.

3.4.4 Design of non-targeting sgRNA as negative controls for CRISPR-Cas9 and CRISPR-Cpf1 screens

As with any experiment, the use of appropriate controls are required to underscore the positive results. For this purpose, the genome-wide Cas9 and Cpf1 sgRNA libraries also include upto 1% of non-targeting sgRNA that function as negative controls. These sgRNA are randomly generated sequences, that theoretically should not be able to direct the endonuclease to cut anywhere in the genome. This section details the algorithm to create negative controls for the Cas9 and Cpf1 sgRNA libraries.

1. The randseq function in MATLAB allows for the generation of a random nucleotide sequence of a specified length. To create appropriate negative control, the generated nucleotide sequence must be blasted against the genome and found to contain mismatches preferably within first few nucleotides of an sgRNA which invariably function as a seed sequence. Since the seed sequence is essential for correct positioning of the nuclease and its subsequent nuclease activity, mismatches within the seed sequence are desired characteristics in a non-targeting sgRNA.

2. In the case of a Cas9-sgRNA, the seed sequence is usually the 10-12 nt on the 3' end of the sequence. To generate the non-targeting sgRNA, we would first need to import the complete sequence of each chromosome of CLIB89 into a variable in MATLAB. As mentioned before this is done using the fastaread command, and stored in the variable named 'Chromosomes'.

3. Following this, we need to continually generate random 20 bp sequences using the randseq command, and determine if the first 10 bp have a match anywhere in the genome. This can be done by, dividing the sequence of each chromosome into 10 bp chunks. Nucleotides 1-10 would occupy a slot; nucleotides 2-11 would occupy another slot, and so on until the end of each chromosome.

4. Finally, the first 10 bp of each randomly generated sgRNA would be compared with these 10 bp slots, and discarded if a match is found. If a match is not found after spanning every slot for each chromosome, then that sgRNA is recorded. When the count of recorded sgRNA reaches 1% of the library size, the iteration is stopped.

5. In the case of a Cpf1-sgRNA, the seed sequence is typically within the first 14 nt, and in some cases even upto the first 17 nt. In our library, non-targeting sgRNA for Cpf1, were determined by continually generating 25 bp sequences using the randseq command and determining if the first 12 bp have a match anywhere in the genome. Similar to the procedure for Cas9, each chromosome would be divided into 12 bp chunks. Nucleotides 1-12 would occupy a slot; nucleotides 2-13 would occupy another slot, and so on until the end of each chromosome.

6. Finally, the first 12 bp of each randomly generated sgRNA would be compared with these 12 bp slots, and discarded if a match is found. If a match is not found after spanning every slot for each chromosome, then that sgRNA is recorded. When the count of recorded sgRNA reaches 1% of the library size, the iteration is stopped.

7. The generated set of non-targeting guides can be verified by mapping this to the genome on a mapping tool such as Bowtie. The mapping statistics on Bowtie should return a 0% exact match.

Table 3.1. Test for uniqueness of Cpf1 sgRNA.

Variable	14 bp Seed	15 bp Seed	16 bp Seed
Sg_Top	97619	97619	97619
Sg_Bottom	93463	93463	93463
Sg_Top_Unique	94350	94893	95136
Sg_Bottom_Unique	89972	90491	90689
% Loss in sgRNA	3.54%	2.98%	2.76%
Final sgRNA List	57771	57883	57856

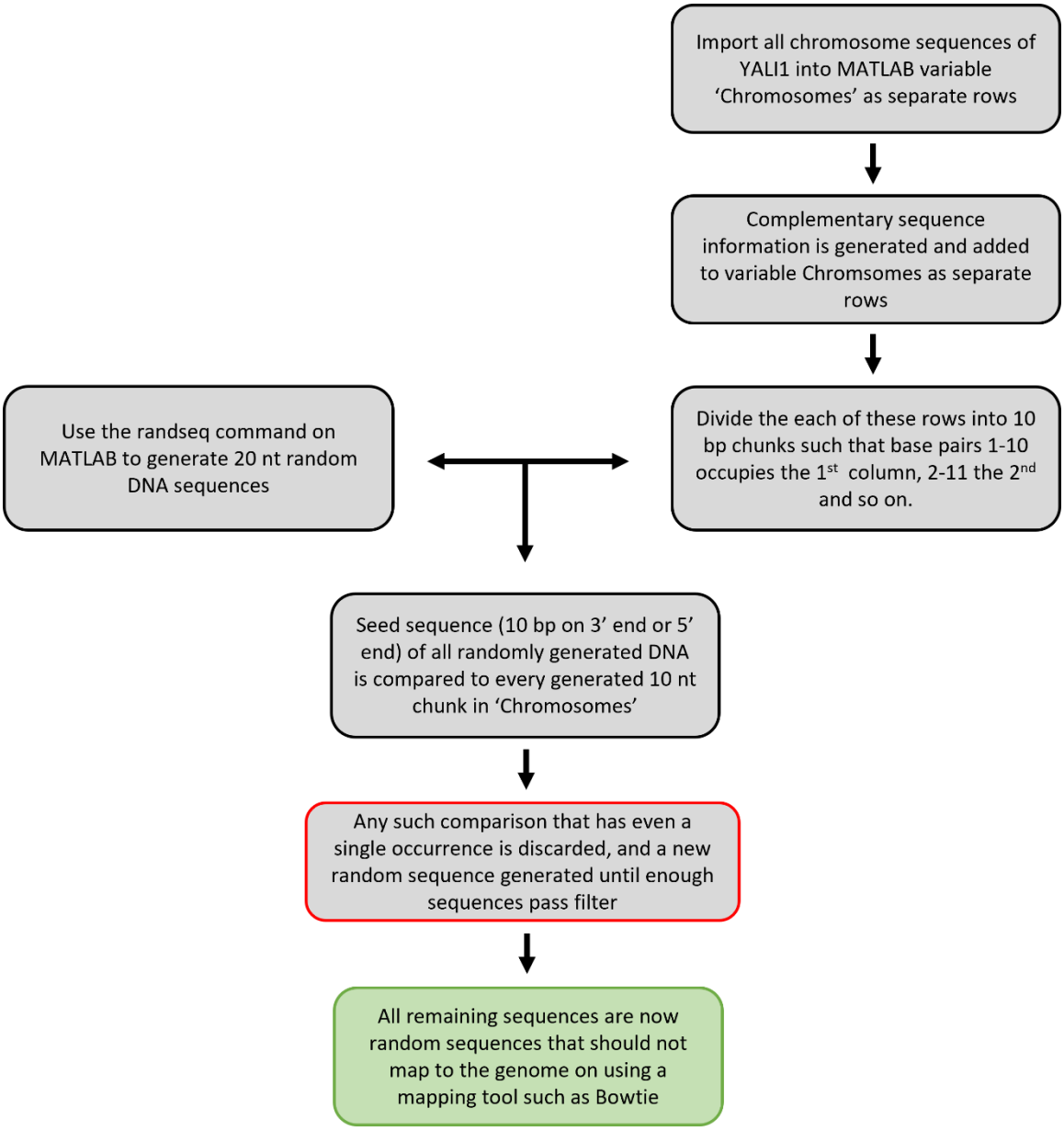

Figure 3.4. Flow diagram for the design nontargting negative controls. The chromosome information is once again stored in a variable and subsequently split into 10 bp chunks each occupying a cell. A seed of 10 bp from randomly generated sgRNA sequences is compared to the 10 bp chromosomal chunks. If there is no match, the sequence is stored as a nontargeting guide. As an added validation, running a genome mapping tool such as Bowtie should also indicate that the generated sequences do not map to the genome

3.5 Notes

1. Criterion for the uniqueness of sgRNA. An important consideration when testing for the uniqueness of an sgRNA is ensuring the uniqueness of its seed sequence. In the case of Cas9-sgRNA, the seed sequence has been defined as the PAM-proximal 10-12 nucleotides located at the 3' end of the 20 bp spacer sequence. Target specificity is strongly influenced by the complementarity between the seed sequence and the genomic target. Mismatches in this seed region severely impede or even completely nullify target DNA binding and nuclease activity of the endonuclease. In the case of Cpf1-sgRNA, it has been found out that mismatches between the spacer and the genomic target at positions 1–8, 10–14, and 17, severely impair cleavage activity of the endonuclease. In the case of certain Cpf1 variants, mismatches within the first 17 nt also showed significant effects on DNA cleavage. With this information in mind, it became a necessity to ensure mismatches within the seed region of the sgRNA to ensure its uniqueness and reduce the possibility of off-target binding and cleavage. As a result, only those Cpf1-sgRNA whose first 14 nt showed at least one mismatch with all other possible Cpf1-sgRNA in the genome were picked. Similarly, only those Cas9-sgRNA that that showed no sequence similarity within the first 12 nt to any other Cas9-sgRNA in the genome was picked. While mismatches within the seed region is the generally accepted criterion for a test for uniqueness, it is always possible to be more stringent by reducing the length of the seed region within which to ensure a mismatch. However, the shorter the length of this region, the more are the sgRNA that will be eliminated from the library. Thus, it is important to strike a

balance in the stringency of the uniqueness criterion to ensure large enough library with theoretically minimal off-target effects. As a test, when we generated the sgRNA for the genome wide Cpf1 sgRNA library, we tested the uniqueness of the sgRNA at 16 bp, 15 bp and 14 bp. As seen from Table 3.1, as we decrease the length of the sequence within which to ensure a mismatch, the more we lose sgRNA. However, at 14 bp we were still able to design a library that should have an 8-fold coverage of over 82% of the genes in *Y. lipolytica*, and at least a 5-fold coverage of over 90% of the genes. Since the test for uniqueness was conducted prior to the selection of sgRNA for the library, the loss in sgRNA due to the uniqueness test was offset if the gene had more than 8 unique sgRNA.

2. MATLAB scripts for Cas9 and Cpf1 sgRNA design. Custom MATLAB scripts that were used for the design of the Cas9 and Cpf1 CRISPR library can be found at the following link:
   https://github.com/ianwheeldon/acCRISPR/tree/main/MATLAB_scripts_genome_wide_CRISPR_screens_Y_lipolytica

3.6 References

1. Warner JR, Reeder PJ, Karimpour-Fard A, Woodruff LBA, Gill RT (2010) Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. Nature Biotechnology 28 (8):856-U138. doi:10.1038/nbt.1653

2. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann M, Fradet-Turcotte A, Sun S, Mero P, Dirks P, Sidhu S, Roth FP, Rissland OS, Durocher D, Angers S, Moffat J (2015) High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell 163 (6). doi:10.1016/j.cell.2015.11.015

3. Sidik SM, Huet D, Ganesan SM, Huynh MH, Wang T, Nasamu AS, Thiru P, Saeij JPJ, Carruthers VB, Niles JC, Lourido S (2016) A Genome-wide CRISPR Screen in Toxoplasma Identifies Essential Apicomplexan Genes. Cell 166 (6):1423-+. doi:10.1016/j.cell.2016.08.019

4. Evers B, Jastrzebski K, Heijmans JPM, Grernrum W, Beijersbergen RL, Bernards R (2016) CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. Nature Biotechnology 34 (6):631-633. doi:10.1038/nbt.3536

5. Bao ZH, HamediRad M, Xue P, Xiao H, Tasan I, Chao R, Liang J, Zhao HM (2018) Genome-scale engineering of Saccharomyces cerevisiae with single-nucleotide precision. Nature Biotechnology 36 (6):505-+. doi:10.1038/nbt.4132

6. Roy KR, Smith JD, Vonesch SC, Lin G, Tu CS, Lederer AR, Chu A, Suresh S, Nguyen M, Horecka J, Tripathi A, Burnett WT, Morgan MA, Schulz J, Orsley KM, Wei W, Aiyar RS, Davis RW, Bankaitis VA, Haber JE, Salit ML, St Onge RP, Steinmetz LM (2018) Multiplexed precision genome editing with trackable genomic barcodes in yeast. Nature Biotechnology 36 (6):512-+. doi:10.1038/nbt.4137

7. Thorwall S, Schwartz C, Chartron JW, Wheeldon I (2020) Stress-tolerant non-conventional microbes enable next-generation chemical biosynthesis. Nat Chem Biol 16 (2):113-121. doi:10.1038/s41589-019-0452-x

8. Löbs A-K, Schwartz C, Wheeldon I (2017) Genome and metabolic engineering in non-conventional yeasts: Current advances and applications. Synthetic and Systems Biotechnology 2 (3):198-207. doi:https://doi.org/10.1016/j.synbio.2017.08.002

9. Blazeck J, Hill A, Liu LQ, Knight R, Miller J, Pan A, Otoupal P, Alper HS (2014) Harnessing Yarrowia lipolytica lipogenesis to create a platform for lipid and biofuel production. Nature Communications 5. doi:10.1038/Ncomms4131

10. Rodriguez GM, Hussain MS, Gambill L, Gao DF, Yaguchi A, Blenner M (2016) Engineering xylose utilization in Yarrowia lipolytica by understanding its cryptic xylose pathway. Biotechnology for Biofuels 9. doi:10.1186/s13068-016-0562-6

11. Yaguchi A, Spagnuolo M, Blenner M (2018) Engineering yeast for utilization of alternative feedstocks. Current Opinion in Biotechnology 53:122-129. doi:10.1016/j.copbio.2017.12.003

12. Schwartz C, Frogue K, Misa J, Wheeldon I (2017) Host and Pathway Engineering for Enhanced Lycopene Biosynthesis in Yarrowia lipolytica. Front Microbiol 8. doi:10.3389/fmicb.2017.02233

13. Qiao K, Imam Abidi SH, Liu H, Zhang H, Chakraborty S, Watson N, Kumaran Ajikumar P, Stephanopoulos G (2015) Engineering lipid overproduction in the oleaginous yeast Yarrowia lipolytica. Metab Eng 29:56-65. doi:10.1016/j.ymben.2015.02.005

14. Xue ZX, Sharpe PL, Hong SP, Yadav NS, Xie DM, Short DR, Damude HG, Rupert RA, Seip JE, Wang J, Pollak DW, Bostick MW, Bosak MD, Macool DJ, Hollerbach DH, Zhang HX, Arcilla DM, Bledsoe SA, Croker K, McCord EF, Tyreus BD, Jackson EN, Zhu Q (2013) Production of omega-3 eicosapentaenoic acid by metabolic engineering of Yarrowia lipolytica. Nature Biotechnology 31 (8):734-+

15. Markham KA, Palmer CM, Chwatko M, Wagner JM, Murray C, Vazquez S, Swaminathan A, Chakravarty I, Lynd NA, Alper HS (2018) Rewiring Yarrowia lipolytica toward triacetic acid lactone for materials generation. Proceedings of the National Academy of Sciences of the United States of America 115 (9):2096-2101. doi:10.1073/pnas.1721203115

16. Schwartz C, Curtis N, Lobs AK, Wheeldon I (2018) Multiplexed CRISPR Activation of Cryptic Sugar Metabolism Enables Yarrowia Lipolytica Growth on Cellobiose. Biotechnol J 13 (9). doi:10.1002/biot.201700584

17. Schwartz C, Frogue K, Ramesh A, Misa J, Wheeldon I (2017) CRISPRi repression of nonhomologous end-joining for enhanced genome engineering via homologous recombination in Yarrowia lipolytica. Biotechnology and Bioengineering 114 (12):2896-2906. doi:10.1002/bit.26404

18. Schwartz C, Shabbir-Hussain M, Frogue K, Blenner M, Wheeldon I (2017) Standardized Markerless Gene Integration for Pathway Engineering in Yarrowia lipolytica. Acs Synth Biol 6 (3):402-409. doi:10.1021/acssynbio.6b00285

19. Schwartz CM, Hussain MS, Blenner M, Wheeldon I (2016) Synthetic RNA Polymerase III Promoters Facilitate High-Efficiency CRISPR-Cas9-Mediated Genome Editing in Yarrowia lipolytica. Acs Synth Biol 5 (4):356-359. doi:10.1021/acssynbio.5b00162

20. Schwartz C, Cheng J-F, Evans R, Schwartz CA, Wagner JM, Anglin S, Beitz A, Pan W, Lonardi S, Blenner M, Alper HS, Yoshikuni Y, Wheeldon I (2019) Validating genome-wide CRISPR-Cas9 function improves screening in the oleaginous yeast Yarrowia lipolytica. Metab Eng 55:102-110. doi:10.1016/j.ymben.2019.06.007

21. Ramesh A, Ong T, Garcia JA, Adams J, Wheeldon I (2020) Guide RNA Engineering Enables Dual Purpose CRISPR-Cpf1 for Simultaneous Gene Editing and Gene Regulation in Yarrowia lipolytica. Acs Synth Biol 9 (4):967-971. doi:10.1021/acssynbio.9b00498

22. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J, Root DE (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature Biotechnology 34 (2):184-+. doi:10.1038/nbt.3437

23. Chuai GH, Ma HH, Yan JF, Chen M, Hong NF, Xue DY, Zhou C, Zhu CY, Chen K, Duan B, Gu F, Qu S, Huang DS, Wei J, Liu Q (2018) DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biology 19. doi:10.1186/s13059-018-1459-4

24. Blazeck J, Liu LQ, Redden H, Alper H (2011) Tuning Gene Expression in Yarrowia lipolytica by a Hybrid Promoter Approach. Appl Environ Microbiol 77 (22):7905-7914. doi:10.1128/Aem.05763-11

25. Blazeck J, Reed B, Garg R, Gerstner R, Pan A, Agarwala V, Alper HS (2013) Generalizing a hybrid synthetic promoter approach in Yarrowia lipolytica. Appl Microbiol Biot 97 (7):3037-3052

26. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL, Xavier RJ, Root DE (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nature Biotechnology 32 (12):1262-U1130. doi:10.1038/nbt.3026

Chapter 4: Genome-wide functional screens enable the prediction of high activity CRISPR-Cas9 and -Cas12a guides in *Yarrowia lipolytica*

4.1 Abstract

Genome-wide functional genetic screens have been successful in discovering genotype-phenotype relationships and in engineering new phenotypes. While broadly applied in mammalian cell lines and in *E. coli*, use in non-conventional microorganisms has been limited, in part, due to the inability to accurately design high activity CRISPR guides in such species. Here, we develop an experimental-computational approach to sgRNA design that is specific to an organism of choice, in this case the oleaginous yeast *Yarrowia lipolytica*. A negative selection screen in the absence of non-homologous end-joining, the dominant DNA repair mechanism, was used to generate single guide RNA (sgRNA) activity profiles for both SpCas9 and LbCas12a. This genome-wide data served as input to a deep learning algorithm, DeepGuide, that is able to accurately predict guide activity. DeepGuide uses unsupervised learning to obtain a compressed representation of the genome, followed by supervised learning to map sgRNA sequence, genomic context, and epigenetic features with guide activity. Experimental validation, both genome-wide and with a subset of selected genes, confirms DeepGuide's ability to accurately predict high activity sgRNAs. DeepGuide provides an organism specific predictor of CRISPR guide activity that with retraining could be applied to other non-conventional microbes.

This chapter previously appeared as an article in *Nature Communications*. The original citation is as follows: Baisya, D., Ramesh, A., Schwartz, C., Lonardi, S., & Wheeldon, I. (2022). Genome-wide functional screens enable the prediction of high activity CRISPR-Cas9 and-Cas12a guides in Yarrowia lipolytica. *Nature communications*, *13*(1), 1-10.

4.2 Introduction

Class II CRISPR endonucleases such as Cas9 and Cas12a are now widely used for targeted genome editing and in functional genomics screens. These multi-domain proteins function by forming a ribonucleoprotein complex of a CRISPR RNA (crRNA or spacer) and a structural component that enables complexation of the crRNA with the CRISPR associated endonuclease (*i.e.*, Cas9 or Cas12a) [1,2]. Targeting is achieved by the complementarity of the crRNA to a desired genomic locus, which must be adjacent to a protospacer adjacent motif (PAM) to activate endonuclease function. When this targeting occurs, active Cas9 or Cas12a can create a loss of function mutation as an endonuclease induced double stranded break in the genome is repaired by native non-homologous end joining (NHEJ) or by homologous recombination (HR) in the presence of a repair template [3,4]. Gene regulation is also possible with Cas activity disabled, by targeting repressor or activation domains to the promoter region of the gene of interest [5]. Such editing and regulation can be accomplished individually [6], in multiplexed format [7] or with pooled libraries of gRNAs that target every gene in a genome [8]. The development of these systems has not only enabled genetic studies in model cell lines and microbes, but have also eased the burden of developing targeted genome editing tools in many non-model or non-conventional organisms [9–14].

The successful application of CRISPR systems is largely dependent on the efficacy of the sgRNA, and while a number of design tools have been developed, accurate predictions across species and across different Cas endonucleases is not yet possible. A central challenge is that the vast majority of predictive algorithms are trained on data generated

from a limited number of species, most commonly human and murine cell lines or *E. coli*. In addition, most screens to date that correlate sgRNA sequence with activity have been conducted with Cas9 or Cas9 variants, with only a limited number of such screens for Cas12a (Cpf1) or other Cas proteins. A recent meta-analysis of CRISPR-Cas9 screens suggests that the lack of cross-species predictive power comes from variation in genomic context; a strong correlation between sgRNA features and guide activity for the target species were not able to predict guide activity when applied to other species [15]. We have also observed this in our own work, where genome-wide sgRNA activity profiles in the oleaginous yeast *Yarrowia lipolytica* showed poor correlation with activity predicted by a number of commonly-used guide design tools trained on data generated from other species[8].

Here, we developed a deep learning-based guide design algorithm called DeepGuide that is capable of accurately predicting *Streptococcus pyogenes* Cas9 and *Lachnospiraceae bacterium* Cas12a sgRNA activity in *Y. lipolytica*. We focused our efforts on this non-conventional yeast because it has value as an industrial 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 [16–21]. Similar to many other eukaryotes, DNA repair in *Yarrowia* is dominated by NHEJ [22]. We exploit this trait to perform negative selection CRISPR screens in the absence of NHEJ repair where double stranded breaks in the genome lead to cell death or a significant impairment to cell fitness [8,23]. Such screens enable the quantification of a cutting score (CS), a measure of activity, for every plasmid expressed sgRNA in the library, thus creating a large data set correlating sgRNA activity

to guide sequence, genomic context, and other genomic and epigenetic features. This work generates a dataset for Cas12a and also uses Cas9 genome-wide CS profiles generated in a previous work [8] to create a large, *Y. lipolytica* specific training set to understand and predict guide activity for CRISPR studies in this yeast.

DeepGuide utilizes a deep learning framework based on a convolutional neural network (CNN), that builds on existing sgRNA activity prediction tools such as DeepCRISPR [24] and Seq-deepCpf1 [25]. Unsupervised learning was achieved using a convolutional autoencoder (CAE) in a pretraining step to learn the representation of the sgRNA landscape within the genomic context of *Y. lipolytica.* This was followed by supervised learning on a CNN using sequence and a CS value for each sgRNA sequence within the Cas9 and Cas12a datasets, and related chromatin accessibility information for the target site of each sgRNA. Lastly, the predictions of the model were cross-validated to obtain correlations between observed and predicted CS values. Activity of predicted guides was also independently validated by targeting a set of genes whose null mutants generated easily screenable phenotypes. DeepGuide outperformed existing guide activity prediction tools on the *Y. lipolytica* datasets and predicted 20 nt Cas9 sgRNA with an NGG PAM, as well as 25 nt Cas12a sgRNA with a TTTV PAM, with high accuracy.

4.3 Results

4.3.1 Library design and generating genome-wide CS profiles

To generate *Y. lipolytica* CS profiles for CRISPR-Cas9 and -Cas12a, we designed plasmid-based sgRNA libraries with 6-fold and 8-fold redundancy for every protein-coding gene in the *Y. lipolytica* genome. The Cas9 library targeted 7,854 out of 7,919 protein-coding genes annotated in the CLIB89 strain (parent strain of PO1f) of *Y. lipolytica* [26], while the more restrictive PAM sequence of Cas12a (TTTV for Cas12a vs. NGG for Cas9) resulted in a library targeting only 7,801 protein coding genes. Gene coverage of the library as well as distributions of the guides within each library after plasmid construction are shown in Figure S4.1. Libraries were designed using two distinct approaches: a strategy biased towards active guides for Cas9, and an unbiased strategy for Cas12a. For the Cas9 library, we used the first iteration of sgRNA Designer [27] to rank all possible Cas9 guides in *Y. lipolytica* and selected the top six scoring guides for every targeted gene (Note: experimental analysis of this library was previously accomplished, including CS profiling, and negative and positive selection screens [8]. Here, we re-analyze this data and use it as training and validation sets for DeepGuide). For the Cas12a library, sgRNAs were selected at random starting from the 5' end of each gene. With the exception of ensuring that the sgRNAs would have minimal or no off-target effects, no additional criteria were used to design the library. We used only minimal design criteria so that a significant portion of the library would contain poorly active or inactive guides. This unbiased Cas12a library was expected to provide a more informative training set for DeepGuide due to the presence of a higher proportion of "negative" training examples.

The workflow to generate the CS profiles, along with the distributions for both Cas9 and Cas12a are shown in Figure 4.1, with replicate correlations shown in Supplementary Figure 4.2 and Supplementary Table S4.1. The CS value for each guide is defined as the $\log_2$ ratio of normalized sgRNA abundance in a NHEJ-deficient strain, to that in a strain both deficient in NHEJ and expressing Cas9/12a (Supplementary Files 4.1 and 4.2). The lack of Cas activity removes a pressure for selection and therefore sgRNA abundance in the control strain was expected to remain relatively constant over the course of the growth screen. Cas9/12a induced double stranded breaks in a strain deficient in NHEJ causes cell death or significantly impairs growth, thus linking sgRNA abundance (as measured by next generation sequencing of the recovered sgRNA expression plasmids) to Cas9/12a activity, where high positive CS values indicate high activity guides and negative CS values indicated inactive or poorly active guides.

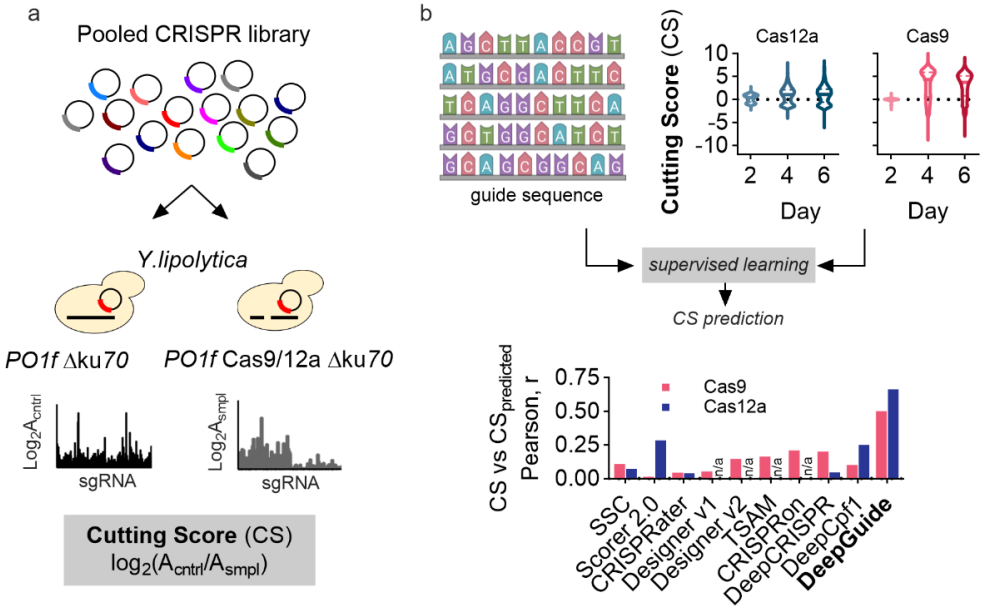

Figure 4.1. Generating genome-wide CRISPR-Cas9 and -Cas12a guide activity scores as input to machine learning algorithms for guide activity prediction. (a) Pooled libraries of single guide RNAs (sgRNAs) for *Streptococcus pyogenes* Cas9 and for *Lachnospiraceae bacterium* Cas12a were transformed into *Y. lipolytica* strains with non-homologous end-joining (NHEJ) DNA repair disabled by disruption of KU70. The sample strain (smpl) expresses Cas9 or Cas12a, while the control strain (cntrl) does not. The Cas12a screens were conducted for this work, while the Cas9 screens were previously reported in ref. [8]. A double stranded CRISPR cut to the genome in the absence of KU70 function leads to cell death (or a dramatic reduction in cell growth), thus enabling the quantification of guide activity through a cutting score (CS) defined as the $\log_2$ fold change of normalized guide abundance in the control vs. the sample determined by next generation sequencing. (b) Genome-wide CS and sgRNA sequence are used as inputs to the convolutional autoencoder (CAE)-based learning method, DeepGuide, to predict sgRNA CS. DeepGuide prediction of Cas9 guides also used as input a normalized score for nucleosome occupancy across the genome [46]. The performance of established CRISPR guide prediction algorithms, including Spacer Scoring for CRISPR (SSC) [29], sgRNA Scorer 2.0 (Scorer 2.0) [30], CRISPRater [28], Designer v1 and v2 [27,31], TSAM [32], CRISPRon [33], DeepCRISPR [24], and Seq-deepCpf1 [25], are shown as a comparison to DeepGuide. The graph shows the Pearson correlation coefficient between CS and the predicted CS for each method. DeepGuide was trained on Cas9 and Cas12a genome-wide CS, the corresponding sgRNA sequence, and genomic context, while all other algorithms used sgRNA sequence (and when appropriate, genomic context) as inputs.

With CS profiles for both Cas9 and Cas12a in-hand, we set out to determine if a number of commonly used guide prediction methods could capture our experimentally determined CS profiles. Learning-based models that use only the sgRNA sequence as input, including CRISPRater [28], SSC [29], and sgRNA Scorer [30] were partially able to capture CS across the genome with SSC exhibiting the highest Pearson coefficient for Cas9 (r = 0.11) and sgRNA Scorer the highest for Cas12a (r = 0.28). sgRNA Designer [27,31] and TSAM [32] take as input the guide sequence and the genomic context immediately surrounding it, but were also not able to accurately capture experimentally determined CS values in *Y. lipolytica*. TSAM performed the best of these (including both versions of sgRNA Designer [27,31]), achieving a Pearson coefficient of r = 0.16 for Cas9. These three algorithms are not designed for Cas12a guide prediction, as such were not able to predict Cas12a CS in *Y. lipolytica*. Lastly, three neural network-based approaches, Seq-deepCpf1 [25], DeepCRISPR [24], and CRISPRon [33], were also only partially aligned with CS; Seq-deepCpf1 fared the best at predicting Cas12a CS (r = 0.25), while CRISPRon was best at predicting Cas9 activity (r = 0.21). DeepGuide, our CAE/CNN-based approach, achieved Pearson coefficients of 0.5 and 0.66 for Cas9 and Cas12a CS values, respectively. We note here that in the case of Cas9, nucleosome occupancy was also used as input to the predictive algorithm; details of this and DeepGuide optimization are discussed in the following subsections.

The comparison of existing methods to DeepGuide were accomplished using CS values after four days of cell growth. CS distributions determined after two, four and six

days are shown in Figure 4.2. After only two days of culture, CS values remained close to zero indicating minimal guide activity (at day 2, $CS_{Cas9,avg}$= -0.01±0.21, $CS_{Cas12a,avg}$ 0.22±0.83). At the end of the second day of growth post-transformation, the sample and control strains reached confluency for the first time and were subcultured to continue the growth screen at this time point as well as after reaching confluency for a second time four days into the screen. We elected to use day 4 data for further analysis because the observed CS profiles remained relatively unchanged from day 4 to day 6, suggesting that the majority of sgRNA activity and the resulting phenotypic effect had occurred by day 4. Both libraries also included a population of non-targeting sgRNAs, constituting ~1.5% of each library, that functioned as negative controls. For both Cas12a and Cas9 the average CS for the negative control populations were in the -1.0 to -3.0 range (across all days) and were represented by normal distributions around -1.56 for Cas12a (day 4) and -3.0 9 for Cas9 (day 4).

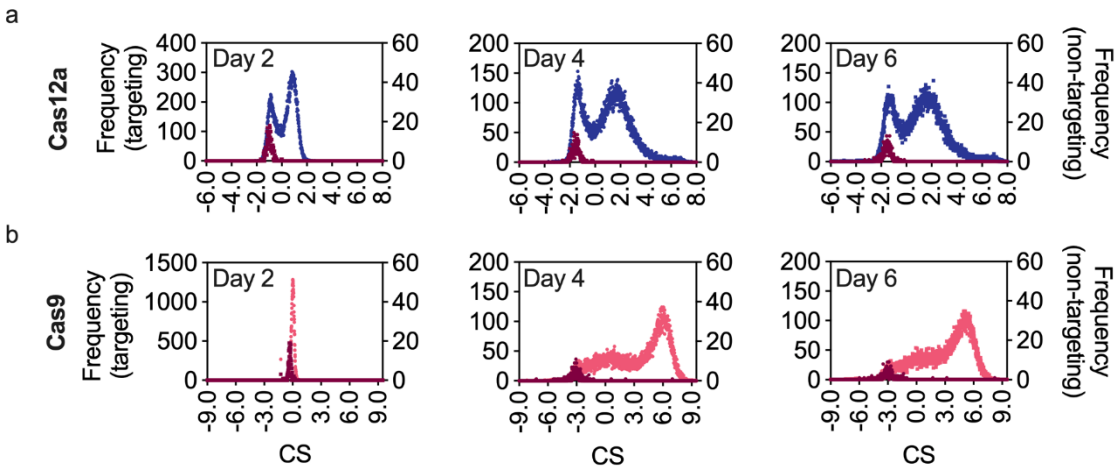

Figure 4.2. CRISPR-Cas12a and -Cas9 cutting score (CS) distributions in *Yarrowia lipolytica*. CS distributions were calculated across three separate days after subculturing transformants twice when they reached confluency. Blue and Pink distributions plotted on the left y-axis show CS values of Cas12a and Cas9 libraries, while the dark red data plotted with the right y-axis depicts the non-cutting control population, constituting ~1% of the respective library. The higher the value of CS, the better the cutting activity of the sgRNA. (a) Histogram of CS values in Cas12a library. (b) Histogram of CS values in Cas9 library. The CS values at Day 4 for both Cas9 and Cas12a were carried forward for further analysis.

4.3.2 DeepGuide architecture and training

DeepGuide consists of three interconnected neural networks, namely a convolutional autoencoder (CAE), a convolutional fully-connected neural network and a small fully-connected network that is used to capture additional epigenetic features (in our case, nucleosome occupancy data; Figure 4.3). The convolutional autoencoder takes as input all the *k*-mers from the genome of interest and builds a compressed representation (in the form of internal weights in the encoder) of the genomic background distribution. The second network is composed of an encoder followed by a fully connected neural network (see Supplementary Table S4.2 for the list of layers). The encoder matches the structure of the encoder in the CAE, and its weights are first initialized from the CAE pre-training step. The fully connected neural network is composed of one flattening layer, three fully connected layers, one concatenation layer, and one output layer (see Supplementary Table S4.3 for the list of layers). The entire second network (including the encoder) is trained via back-propagation from input pairs of sgRNA sequences and their corresponding CS values. The nucleosome data is fed into the third fully-connected neural network. One-dimensional occupancy data is expanded into a multi-dimensional real vector using a fully connected layer. The output layer of this third network is finally combined using element-wise multiplication with the output layer of the second network to generate CS predictions that account for the sgRNA sequence, genomic context, and nucleosome occupancy. Additional details with respect to these architectures and their training are provided in the Material and Methods section.

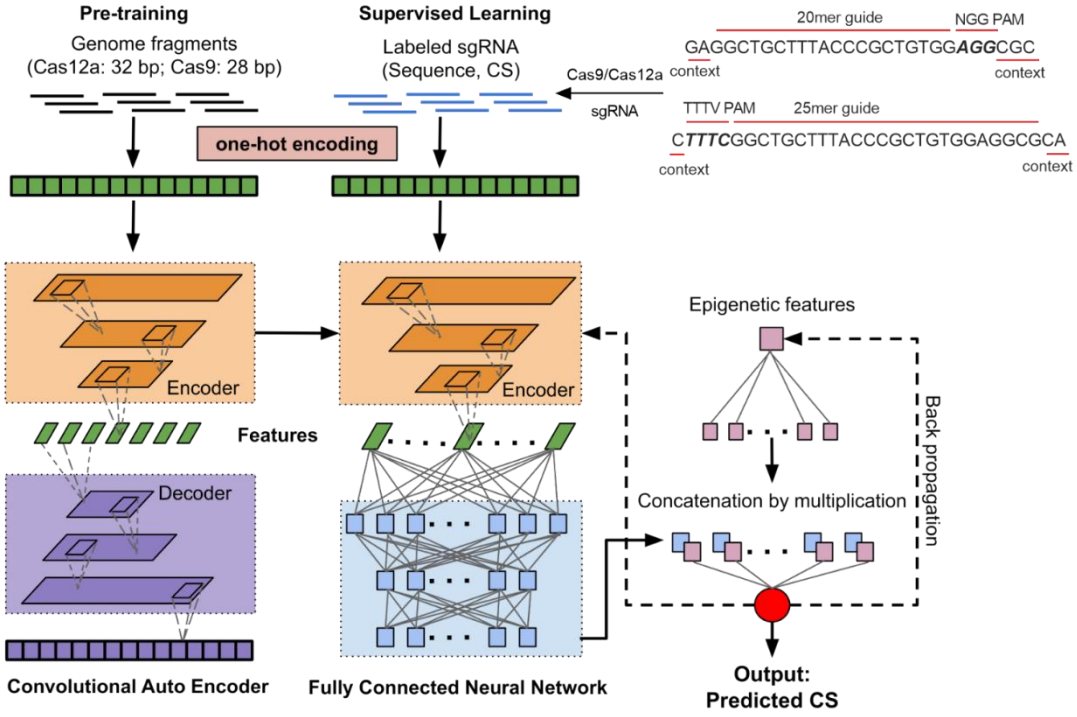

Figure 4.3. The architecture of DeepGuide. First, the entire *Y. lipolytica* PO1f genome was fragmented into sgRNA sized chunks (using a sliding window of 20 bp for Cas9 and 25 bp for Cas12a). Unsupervised pre-training was carried out on these unlabeled fragments using a convolutional autoencoder (left). The internal weights from the autoencoder were used to initialize a fully connected convolutional neural network (center). Labeled sgRNA (*i.e.*, sequence and associated cutting score) were used as inputs for back-propagation learning on the fully connected neural network. See Tables S2-3 for a description of the layers.

4.3.3 DeepGuide optimization

The choice of a CAE combined with a fully-connected CNN was motivated by the results of a five-fold cross-validation performance evaluation among various machine learning methods (Figure 4.4a). The compared methods include support vector machines (SVM); gradient boosting (GBR), logistic and linear regression; random forests (RF); and, a fully connected neural network (FCNN). As judged by Pearson and Spearman

correlations of the predicted CS and experimentally determined CS, the core CAE/CNN architecture of DeepGuide performed better than all other tested methods. For Cas12a, DeepGuide achieved a Pearson r-value of 0.66 and a Spearman r-value of 0.66, while for Cas9 Pearson and Spearman values were 0.43 and 0.37, respectively. The inclusion of nucleosome occupancy data improved Cas9 prediction accuracy, increasing the Pearson and Spearman r-values to 0.50 and 0.43, respectively. This effect is in agreement with observations of nucleosome inhibition of Cas9/12a targeting *in vitro* and *in viv*o [34–37]. A similar nucleosome occupancy effect on DeepGuide's ability to predict Cas12a CS values, however, was not observed here.

One important question about the performance of any machine learning method relates to the size of the training set, that is, how much data is necessary to obtain the best predictions and what performance penalty is incurred when the training dataset size is limited. Figure 4.4b shows the Pearson and Spearman correlations for DeepGuide as the size of the dataset increases, up to the full-size dataset correlating sgRNA sequence to experimentally determined CS. This analysis shows that (i) DeepGuide's performance improves as the size of the training set increases for both Cas12a and Cas9, and (ii) the performance for Cas9 plateaus as dataset size increases above ~30,000 examples. While the performance curve for Cas12a appears to indicate that a larger dataset could potentially improve performance, the trend still shows that the correlations start to plateau above a training set size of ~30,000.

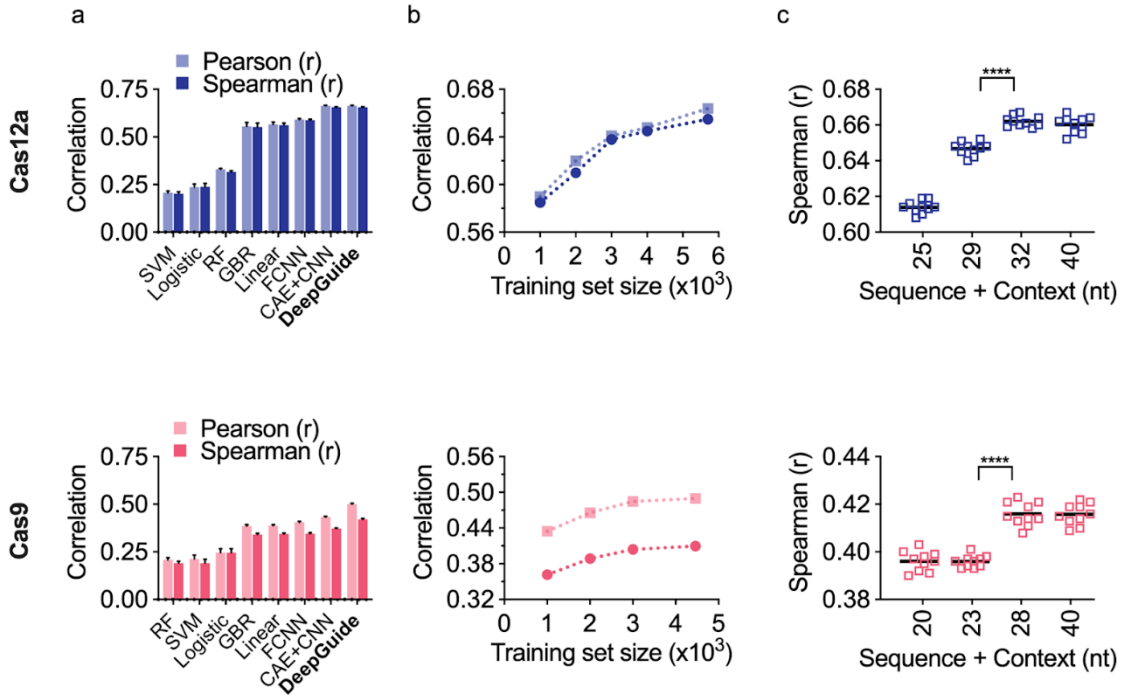

Figure 4.4. Design and parameter optimization for DeepGuide on the Cas12a (top) and Cas9 (bottom) datasets. (a) Evaluation of DeepGuide in a cross-validation analysis with several machine learning (ML) methods, including random forest (RF), support vector machines (SVM), logistic regression (Logistic), gradient boosting regression (GBR), linear regression (Linear), fully-connected neural networks (FCNN), and the core architecture of DeepGuide, a combination of a convolutional autoencoder and a convolutional fully-connected neural network (CAE+CNN). In addition to interconnected CAE and CNN, the final architecture of DeepGuide also includes a third fully connected network to account for nucleosome occupancy. Error bars indicate standard deviation over five independent cross-validation experiments. (b) The dependency of DeepGuide's performance as a function of the training set size with smaller datasets produced by downsampling. (c) The dependency of DeepGuide's performance as function on the length of the context sequence around the sgRNA (ten-fold cross validation). One-way ANOVA indicates that sequence length has a significant effect (**** $p<0.0001$) for both Cas12a and Cas9. Tukey's multiple comparison post-hoc analysis indicates that for Cas12a the Spearman values for all sequence lengths, with exception of 32 vs. 40 bp ($p=0.708$), are significantly different ($p<0.0001$). For Cas9, Tukey's multiple comparisons indicates that all values are significantly different ($p<0.0001$) with the exceptions of 20 vs 23 ($p=0.9995$) and 28 vs 40 bp ($p>0.9999$).

DeepGuide's hyperparameters (*e.g.*, number of hidden layers, number of neurons in each layer, type of activation function, learning rate, etc.) were also optimized using cross-validation. To determine the optimal number of hidden layers in the fully connected neural network downstream of the encoder, we carried out an ablation analysis as described in the next section. Among the input hyperparameters, the length of the context around the sgRNA significantly affected prediction performance. Observe that sequence lengths from 32-40 bp resulted in the best performance for Cas12a; 32 bp was selected because it produced a model with a smaller number of parameters, thus reducing the possibility of overfitting (Figure 4.4c). Similarly, for Cas9 28 bp was selected from a range of 20-40 bp as it produced the best prediction performance.

4.3.4 Ablation analysis of DeepGuide

To understand how pre-training and the number of fully connected layers (downstream of the encoder in the second network) affects DeepGuide's performance, an ablation analysis was performed. First, as a "sanity" check, the encoder alone (*i.e.*, no fully connected layers, but a flatten layer to get a single output) was tested on Cas12a and Cas9 data without any training or pre-training (*i.e.*, using random weights). Observe in the first row of Table 4.1 (also see Tables S4.4 and S4.5) that Spearman and Pearson are essentially zero, as expected. Second, random weights were used for the encoder, then back propagation was run on the flatten layer. Observe in the second row that training just one layer resulted in a significant jump in prediction performance on both data sets. In rows 3-7, the weights of the encoder were initialized from the pre-training step (CAE) and back-propagation was run exclusively on the fully connected layers downstream of the encoder,

that is by freezing the pre-trained weights of the encoder. Under these conditions, the performance was measured by incrementally adding one fully connected layer at the time. By comparing row 2 to row 3, observe that pre-training improves the performance for both Cas12a and Cas9, but more so for Cas12a. Also observe in rows 3-7 that the best performance on the Cas12a data set is obtained when the second network includes only one fully connected layer (fc_8). Similarly, rows 3-7 show that none of the fully connected layers (fc_8, fc_9, fc_{10}) help to improve the performance on the Cas9 data set. However, a significant performance improvement was gained for Cas9 by introducing the multiplication layer ($mult_{11}$), which combines the nucleosome occupancy.

If backpropagation is allowed to fine tune the weights of the encoder, the overall performance improvement is striking (*i.e.*, compare rows 3-7 with rows 8-12). Observe that in the case of Cas12a, one additional fully connected layer (fc_9) helps the performance but adding more is detrimental. As a result of this ablation analysis, the third fully connected layer (fc_{10}) and the multiplication layer ($mult_{11}$) were removed from DeepGuide's architecture for Cas12a guides.

On Cas9, observe in Table 4.1 that adding one fully connected layer (fc_8) improves the performance, but the biggest improvement is due to the multiplication layer ($mult_{11}$) that incorporates the nucleosome occupancy data. As a result of this ablation analysis, the second and third fully connected layers (fc_9 and fc_{10}) were removed from DeepGuide's architecture for Cas9 guides.

Table 4.1. DeepGuide ablation analysis. Row 1 (green) shows the performance of the encoder (followed by a flatten layer) using random weights (no pre-training or backpropagation); row 2 (purple) show the performance of the encoder (followed by a flatten layer) using random weights and then performing back-propagation only on the flatten layer; rows 3-7 (blue) show the performance after pre-training the encoder and then running back-propagation only layers downstream of the encoder; rows 8-12 (pink) show the performance after pre-training and then running back-propagation on the whole network (including the encoder); correlation coefficients in bold corresponds to the best performance; fc = fully connected layer; pool = pooling layer; flatten = flatten layer; mult = multiplication layer (see Tables S3 for the list of layers).

Row	Training	Layer	Pearson, r	
			Cas12a	Cas9
1	random weights	encoder⇨flatten$_7$	0.070	0.003
2	back prop-flatten$_7$	encoder⇨flatten$_7$	0.455	0.312
3	pretrain + back prop flatten$_7$⇨fc$_{8-10}$⇨mult$_{11}$	encoder⇨flatten$_7$	0.532	0.353
4		encoder⇨flatten$_7$⇨fc$_8$	**0.534**	0.310
5		encoder⇨flatten$_7$⇨fc$_8$⇨fc$_9$	0.517	0.291
6		encoder⇨flatten$_7$⇨fc$_8$⇨fc$_9$⇨fc$_{10}$	0.514	0.305
7		encoder⇨flatten$_7$⇨fc$_8$⇨fc$_9$⇨fc$_{10}$⇨mult$_{11}$	0.514	**0.388**
8	pretrain + back prop-all	encoder⇨flatten$_7$	0.641	0.409
9		encoder⇨flatten$_7$⇨fc$_8$	0.658	0.424
10		encoder⇨flatten$_7$⇨fc$_8$⇨fc$_9$	**0.664**	0.414
11		encoder⇨flatten$_7$⇨fc$_8$⇨fc$_9$⇨fc$_{10}$	0.664	0.414
12		encoder⇨flatten$_7$⇨fc$_8$⇨fc$_9$⇨fc$_{10}$⇨mult$_{11}$	0.664	**0.501**

4.3.5 External and internal validation of DeepGuide

Given the optimized DeepGuide architecture, we next set out to measure its ability to predict Cas9 and Cas12a sgRNA activity as measured in single gene disruption experiments. To do so, we used DeepGuide to predict five high activity and five poor activity Cas9 and Cas12a sgRNAs for four genes whose disruption can be measured with an easily screenable phenotype (Supplementary Figure S4.3). These genes included MFE1,

the knockout of which prevents growth on long chain fatty acids; CAN1, which is involved in resistance to L-canavanine; and MGA1 and RAS2, knockouts of which result in colonies with a smooth appearance due to loss of pseudohyphae formation. Plasmids expressing each of the sgRNAs were individually transformed into *Y. lipolytica* in biological triplicate and screened for the presence or absence of the targeted phenotype. For high activity Cas9 guides, the predicted CS ranged from 4.65 to 5.19, while for Cas12a the CS values of the highest activity guides ranged from 1.09 to 2.08. At the lower end, poor-activity guides ranged from -1.12 to 1.88 for Cas9 and -0.72 to 1.00 for Cas12a. The near overlap of CS values in the low and high predicted activity groups for Cas12a is due to the fact that only 12 TTTV PAM sequences are contained within MGA1, thus providing a limited set to select from. The ten guides that provided the largest range were selected even though two of these had nearly equal predicted CS values (for MGA1, $CS_{predicted}$ = 1.09 was included in the high activity group, while $CS_{predicted}$ = 1.00 was included in the low activity group).

DeepGuide was generally successful in predicting active sgRNAs for both Cas12a and Cas9 but had limited ability to accurately predict low activity guides for Cas9 (Figure 4.5a,b). Seventeen of the twenty Cas12a guides that were predicted to be of high activity, clustered together with a mean disruption efficiency of 77.4% and a $CS_{predicted}$ of 1.67 (Supplementary Figure S4.4). Three guides from the high activity group, $CS_{predicted}$ of 1.91, 1.65, and 1.09, did not cluster well with the others and exhibited disruption efficiencies of 24.6%, 19.1%, and 4.8%, respectively. Predicting the lower end of the activity scale was also successful for Cas12a where 20 of 20 guides clustered together with an average disruption efficiency of 12.1% and a $CS_{predicted}$ of 0.16. Predictions for highly active Cas9

guides were also accurate; 18 of 20 sgRNAs clustered together with an average disruption efficiency and $CS_{predicted}$ of 69.8% and 4.91, respectively. However, DeepGuide performed poorly in predicting low or inactive guides for Cas9. Only 4 of 20 sgRNAs in the low activity group exhibited disruption efficiencies below 25%, another nine sgRNAs achieved efficiencies between 25% and 50%, while the remaining seven proved to be highly active with disruption efficiencies above 50%. One explanation for the discrepancy in performance between Cas9 and Cas12a is the difference in training sets. The Cas9 library was biased from the outset towards high activity guides, thus limiting the number of poor activity guides to learn from. Conversely, the Cas12a sgRNA library was not biased from the outset and consequently resulted in a CS data set that included a high number of both poorly active and highly active guides.

In addition to the external validation where individual sgRNAs were tested for disruption efficiency, we also sought to evaluate DeepGuide's ability to discriminate between active and inactive sgRNAs as measured in the pooled screens; that is, we compared experimentally determined CS vs. predicted CS in a ROC analysis. To do so, the mean CS values of the high activity clusters for Cas9 and Cas12a were taken as the threshold for binarizing guide activity. Guides with CS > 1.67 for Cas12a and CS > 4.91 for Cas9 were classified as active, and CS values below this threshold were classified as inactive. DeepGuide outperformed all other tools in classifying highly active guides as indicated by an AUROC of 0.77 for Cas12a and 0.73 for Cas9 (Figures 4.5c,d and Supplementary Figure S4.5). It is important to note that when seq-DeepCpf1, a prediction algorithm with similar architecture as DeepGuide, was retrained on the Cas9 and Cas12a

CS profile generated in *Yarrowia*, the AUROC curve improved from 0.61 to 0.72 for Cas12a and 0.58 to 0.69 for Cas9, underscoring the importance of the dataset used for training the machine learning model.

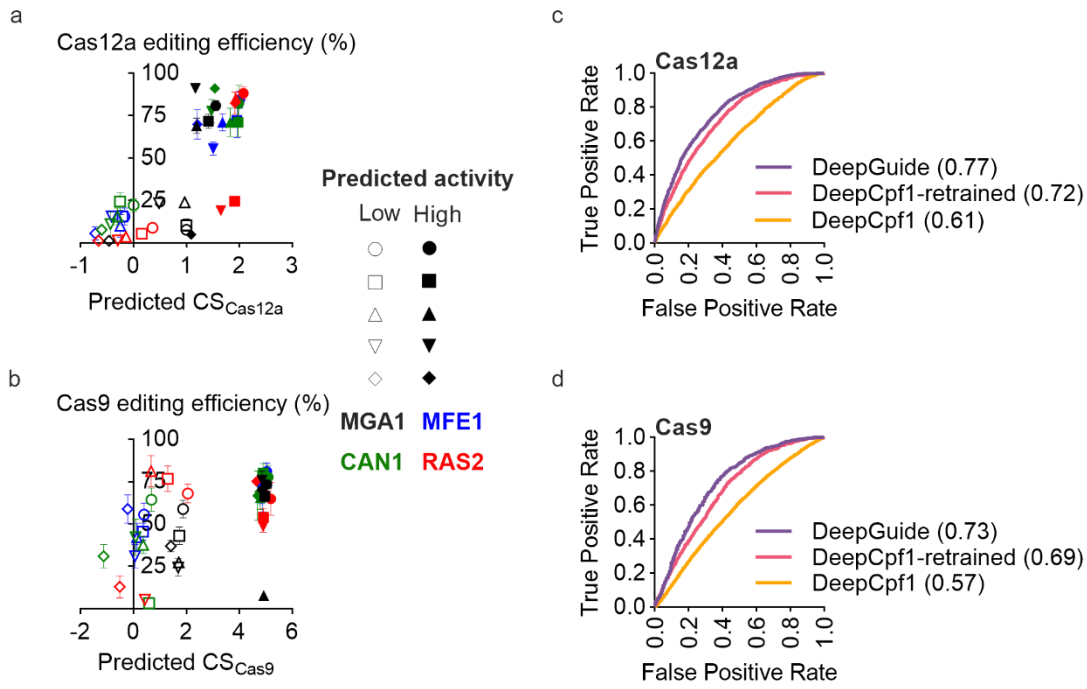

Figure 4.5. External and internal validation of DeepGuide performance. (a) and (b) editing efficiencies of 5 predicted high-activity and 5 predicted low-activity sgRNA for Cas12a and Cas9 in single gene disruption experiments. Genes MGA1, MFE1, CAN1, and RAS2 were picked as their null mutants displayed easily screenable phenotypes. Predicted high-activity sgRNAs clustered together, while low activity sgRNAs clustered at lower editing efficiencies for Cas12a. Data points represent the mean of three biologically independent samples (*n=3*), while the error bars represent the standard deviation. (c) and (d) ROC plots and AUROC values for DeepGuide prediction of high- and low-activity Cas9 and Cas12a sgRNAs.

4.4 Discussion

Current prediction methods have proven effective at designing active CRISPR sgRNAs [24,25,27–33,38,39], but the predictive power is typically limited to the organism from which the training data was generated [8,15]. In this context we created DeepGuide, a machine learning approach to design sgRNA guides based on an organism-specific training set. An evaluation of several machine learning methods and models (see Figure 4.1) allowed us to choose the combination of architectures that would achieve the best predictive performance on our *Y. lipolytica* datasets. When trained on genome-wide CS profiles for both Cas12a and Cas9, DeepGuide accurately designed sgRNA sequences that resulted in high genome editing efficiency (Figure 4.5) and outperformed other methods in predicting Cas9 and Cas12a activity across the genome. Ablation analysis revealed that the organism specific nature of DeepGuide is not solely related to the sgRNA training set but also the genomic context; predictions improved for both Cas9 and Cas12a if DeepGuide's internal weights were initialized via a genome-wide unsupervised learning step on the *Y. lipolytica* genome, rather than being assigned at random (Figure 4.3 and Table 4.1). With retraining, DeepGuide was able to predict guide activity in *E. coli* (see ref. [38]) with good accuracy (see Supplementary Figure S4.7). Given the significant differences in genomic context and methods of generating genome-wide activity profiles, we found that Yarrowia-optimized DeepGuide was not able to accurately predict sgRNA activity in mammalian cells to the same level as the bidirectional long short-term memory neural networks (LSTM) methods that were highly-optimized on such datasets (see ref. [39]; Supplementary Figure S4.7).

While DeepGuide was successful in designing active guides for both Cas12a and Cas9, our analysis and validation experiments revealed significant differences between the two systems. The first was that DeepGuide performed much better on the Cas12a dataset (Cas12a Pearson, r = 0.66 vs. Cas9, r = 0.50), possibly due to the fact that the Cas12a library covers a greater fraction of the total Cas12a PAM sites within the genome (there are 809,401 TTTN PAM sites for Cas12a in *Y. lipolytica* and 2,415,425 Cas9 NGG PAM sites). Library design could also be a driving factor; DeepGuide was not able to accurately predict poor activity guides for Cas9, a result that we ascribe to the low number of 'negative' examples in the biased library designed for Cas9. Lastly, sequence and genomic context were sufficient to drive accurate predictions for Cas12a, but additional contextual information in the form of nucleosome occupancy was necessary to obtain the maximal predictive power for Cas9. The difference in predictive performance between the two systems highlights the importance of having a 'good' training set, in particular for deep learning architectures. A good training set for CRISPR sgRNA prediction should represent high and low activity guides equally, should uniformly sample the entire genome-wide *k*-mer space, should be noise-free (*i.e.*, the guide activity scores should be accurate), and should be sufficiently large (*e.g.*, tens of thousands data points or more).

While this work focuses on the development of DeepGuide for its specific use in *Y. lipolytica*, the same experimental-computational workflow that involves (i) library design, (ii) generating genome-wide guide activity profiles, (iii) predictor design (learning and optimization) and (iv) external validation, can be readily applied to other fungal species, broadly to prokaryotes, and any other organisms in which genome-wide functional screens

can be used to estimate sgRNA activities. Moreover, DeepGuide adds to the growing number of examples in which deep learning is being used to solve complex problems in molecular biology, *e.g.*, the prediction of essential genes [40,41].

4.5 **Methods**

4.5.1 DeepGuide architecture

DeepGuide uses a convolutional autoencoder (CAE) to derive a reduced-dimensionality representation of the underlying distribution of sgRNA sequences in the whole genome. The autoencoder is composed of an encoder (6 layers) and a decoder (6 layers). The objective of the unsupervised training is to infer the internal weight so the input layer to the encoder is as close as possible as the output layer of the decoder. The CAE encoder has two Conv1D layers of 20 filters and 40 filters, respectively, one MaxPooling1D layer, one AveragePooling1D layer and two BatchNormalization layers (see Supplementary Table 2 for the order). A rectified linear activation function (ReLU) is used as activation and the Glorot uniform initializer is used to initialize the convolutional filters. The layer regularizers for the encoder is L2 with value 10E-4. The decoder has the same structure as the encoder but uses UpSampling1D instead of MaxPooling, and UpSampling1D instead of AveragePooling1D. The layer regularizer in the decoder is again L2 with value 10E-4. The loss function for training is the binary cross entropy, and Adam is the optimizer with a learning rate of 10E-3. A batch size of 64 and 200 epochs are used for training (no early stopping).

The encoder in the second network has the same structure of the encoder in the CAE (see Supplementary Table S4.3). The initial configuration of the network downstream of the encoder uses one flatten layer, three fully connected layers (fc8, fc9, fc10) of 80 neurons, 40 neurons and 40 neurons, respectively. The feature map for layer pool6 is 7 x 40 which is 280 dimensional. The feature map for the first fully connected layer (fc8) is 280 x 80 = 22400 dimensional. The feature map for the second and third fully connected layers (fc9 and fc10) are 80 x 40 = 3200 and 40 x 40 = 1600 dimensional, respectively. Layer mult11 is a multiplication layer that combines sequence and nucleosome occupancy features. ReLU is the activation and Glorot uniform initializer is used to initialize the convolutional filters. The second network is trained for 150 epochs using backpropagation; if the value of loss function does not improve for 15 consecutive epochs the training is terminated.

The third fully connected network is used to provide DeepGuide with nucleosome occupancy data. The nucleosome occupancy for each sgRNA is a floating-point value in [0,1]. The third network uses one fully connected layer with 40 units to expand the one-dimensional nucleosome occupancy value to a 40-dimensional vector, to match the dimensionality of the output layer of the second network. Sequence and nucleosome data are merged by performing an element-wise multiplication between the output layer of the second network and the output layer of the third network. When DeepGuide is used in "classification mode" (*i.e.*, binary output) the activation function is sigmoid; when DeepGuide is used in "regression mode" (*i.e.*, cutting score output), the activation function is linear.

Note that following the ablation analysis, only two fully connected layers (and no multiplication layer) are used for Cas12a; similarly, only one fully connected layer connected to the multiplication layer is used for Cas9.

4.5.2 DeepGuide training and pre-training

For the pre-training step of the CAE all *k*-mers from the *Y. lipolytica* genome were extracted using a sliding window of 1 bp. For Cas9 the input length was 28 bp, which includes the length of each possible spacer (20 bp), plus 3 bp for a PAM sequence, and 2 bp up- and downstream for context. For Cas12a, 32-mers were used to account for the 25 bp spacer, a 4 bp PAM, 1 bp of context upstream of the PAM, and 2 bp of context downstream of the spacer (see Figure 4.4b). These unlabeled sgRNA data sets contained over 20 million *k*-mers each. sgRNA sequences were converted into a numerical representation using one-hot encoding, that is, each sgRNA was converted into a 4 x *n* dimensional binary matrix where *n* is the length of the guide.

The training data to DeepGuide consisted of sgRNA sequences, their nucleosome occupancy score, and their CS values. sgRNA sequences were one-hot encoded, while nucleosome occupancy data was processed as explained in the "Nucleosome occupancy analysis" subsection below. CS scores were produced as explained in the "Cutting Score analysis" subsection also provided below.

When the pre-training concluded, the internal weights of the CAE were used to initialize the encoder in the second network. The second network was trained via back-

propagation using either ~45,000 sgRNAs for Cas9 or ~58,000 sgRNA for Cas12a, each with their associated CS value. 60% of these guides were used for training, 20% for validation and 20% for testing. The training step not only allowed the inference of the weights for the fully connected layers downstream of the encoder, but also fine-tuned the weights of the encoder. As explained in the Section "Ablation analysis of DeepGuide" (main text) the pre-training step helped the supervised learning to converge faster and improved the prediction performance.

Supplementary Figure S4.6 illustrates the loss curve for training and validation of the CNN without pre-training and with pre-training as a function on the number of training epochs. Observe that in the CNN without pre-training the difference between training and validation loss function starts increasing after about 20 epochs. In contrast, for the CNN with pre-training the training and validation curves of the loss function are overlapping after about 30 epochs. This indicates that the pre-training prevents the network from overfitting and helps the network to generalize better.

4.5.3 sgRNA library design

Custom Matlab scripts were used to design an LbCas12a sgRNA library with ~8-fold coverage of all protein coding sequences annotated in the *Y. lipolytica* PO1f parent strain genome, CLIB89 [26]. A list of 25 nucleotide (nt) sgRNA with a TTTV (V=A/G/C) PAM were identified in both the top and bottom strand of the coding sequence of each gene (CDS). A second list containing all possible 25nt sgRNAs with a TTTN PAM from the top and bottom strands of all 6 chromosomes in *Y. lipolytica* was also generated and used to

test for sgRNA uniqueness. The uniqueness test was carried out by comparing the first 14nt of each sgRNA in the first list to the first 14nt of every sgRNA in the second list. If a sequence occurred more than once, the sgRNA was identified as non-unique and excluded from consideration. The sgRNAs that passed the test for uniqueness were then picked in an unbiased manner, with even representation from top and bottom strands when possible, starting from the 5' end of the CDS. Six-hundred and fifty-one sgRNAs of random sequence confirmed to not target in the genome were also designed using a similar methodology but with a more stringent criteria for uniqueness (*i.e.*, first 10nt were not found anywhere in the genome). A detailed procedure of sgRNA design for both Cas9 and Cas12a is provided in ref. [42] and additional data on the Cas9 guide design criteria is provided in ref.[8]. Briefly, for Cas9 sgRNAs the first version of sgRNA Designer [27] was used to identify the top predicted guides for every CDS, these guides were filtered for uniqueness, and the top six unique guides were selected.

4.5.3 Microbial strains and culturing

The parent yeast strain used in this study was *Yarrowia lipolytica* PO1f with genotype MatA, *leu2-270*, *ura3-302*, *xpr2-322*, *axp-2*. The PO1f Cas9 and the PO1f Cas12a strains were constructed by integrating UAS1B8-TEF(136)-Cas9-CYCT and UAS1B8-TEF(136)-LbCpf1-CYCT expression cassettes into the A08 locus [43]. The PO1f Cas9 *ku70* and PO1f Cas12a *ku70* strains were constructed by disrupting KU70 using CRISPR-Cas9 as previously described [23]. All strains used in this study are listed in Supplementary Table 6. All plasmid construction and propagation was conducted in *Escherichia coli* TOP10. Cultures were conducted in Luria-Bertani (LB) broth with 100 mg L^{-1} ampicillin at 37 °C

in 14 mL polypropylene tubes, at 225 rpm. Plasmids were isolated from *E. coli* cultures using the Zymo Research Plasmid Miniprep Kit.

4.5.4 Plasmid construction

All plasmids and primers used in this work are listed in Supplementary Table S4.7 and S4.8. To create the LbCas12a sgRNA expression plasmid (pLbCas12ayl), we first added a second direct repeat sequence at the 5' of the polyT terminator in pCpf1_yl (see ref. [44]). This was done to ensure that library sgRNAs could end in one or more thymine residues without being construed as part of the terminator. To make this change, pCpf1_yl was first linearized by digestion with SpeI. Subsequently, primers ExtraDR-F and ExtraDR-R were annealed and this double stranded fragment was used to circularize the vector (NEBuilder® HiFi DNA Assembly) For integrating LbCas12a, pHR_A08_LbCas12a was constructed by digesting pHR_A08_hrGFP (Addgene #84615) with BssHII and NheI, and the LbCas12a fragment was inserted using the New England BioLab (NEB) NEBuilder® HiFi DNA Assembly Master Mix. The LbCas12a fragment was amplified along with the necessary overlaps by PCR using Cpf1-Int-F and Cpf1-Int-R primers from pLbCas12ayl. Successful cloning of the entire fragment was confirmed with sequencing primers A08-Seq-F, A08-Seq-R, Tef-Seq-F, Lb1-R, Lb2-F, Lb3-F, Lb4-F, and Lb5-F. To create the Cas12a sgRNA genome-wide library expression plasmid (pLbCas12ayl-GW) the UAS1B8-TEF- LbCas12a-CYC1 fragment was removed from pLbCas12ayl with the use of XmaI and HindIII restriction enzymes. Subsequently, the primers BRIDGE-F and BRIDGE-R were used to circularize the vector, and the M13 forward primer was used to ensure correct assembly of the construct.

To conduct the validation experiments of predicted CS values by DeepGuide, four genes with easily screenable phenotypes were selected and 10 sgRNA (five highly active and five with poor activity) targeting each of these genes for Cas9 and Cas12a were selected and cloned for individual disruption experiments. All 40 Cas9 sgRNAs with required overlaps for cloning were purchased from a commercial vendor (IDT-DNA) as single stranded primers, and assembled into pCRISPRyl (Addgene #70007) after linearizing the vector with AvrII, using NEBuilder® HiFi DNA Assembly. In a similar manner, the 40 Cas12a sgRNAs with necessary overlaps were cloned into pLbCas12ayl, after linearizing the vector with SpeI. These primers are also included in Supplementary Table S4.8.

4.5.5 sgRNA library cloning

The LbCas12a library targeting the protein coding genes in PO1f were ordered as an oligonucleotide pool from Agilent Technologies Inc. and cloned in-house using the Agilent SureVector CRISPR Library Cloning Kit (Part Number G7556A). The backbone vector (pLbCas12ayl-GW) was first linearized by PCR using the primers InversePCR-F and InversePCR-R, DpnI digested, cleaned up using Beckman AMPure XP SPRI beads, and transformed into *E.coli* TOP10 cells to verify minimal contamination from the circularized plasmid. Library oligos were amplified by PCR using the primers OLS-F and OLS-R for 15 cycles as per vendor instructions using Q5 high fidelity polymerase and cleaned up using the AMPure XP beads. The linearized backbone and the amplicons were combined in 4 replicate reactions of sgRNA library cloning that were carried out as per vendor

instructions and pooled prior to bead cleanup. Two amplification bottles containing 1L of LB media and 3 g of library grade low gelling agarose were prepared, autoclaved, and cooled to 37 °C. Eighteen replicate transformations of the cloned library were conducted using Agilent's ElectroTen-Blue cells (Catalog #200159) via electroporation (0.2 cm cuvette, 2.5 kV, 1 pulse). Cells were recovered and with a 1 hr outgrowth in SOC media at 37 °C (2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, and 20 mM glucose.) The transformed *E. coli* cells were then inoculated into two amplification bottles and grown for 2 days until colonies were visibly suspended in the matrix. Colonies were recovered by centrifugation and subject to a second amplification step by inoculating a 800 mL LB culture. After 4 hr, the cells were collected, and the pooled plasmid library was isolated using the ZymoPURE II Plasmid Gigaprep Kit (Catalog #D4202) yielding ~2.4 mg of plasmid DNA containing the Cas12a sgRNA library. The library was subject to a NextSeq run to test for fold coverage of individual sgRNA and skew.

4.5.6 Yeast transformation and screening

Transformation of *Y. lipolytica* with the sgRNA plasmid library was done using a previously described method with slight modifications [8]. Briefly, 3 mL of YPD was inoculated with a single colony of the strain of interest and grown in a 14 mL tube at 30 °C with shaking at 200 RPM for 22-24 hours (final OD ~30). Cells were pelleted by centrifugation (6,300g) and washed with 1.2 mL of transformation buffer (0.1 M LiAc, 10 mM Tris (pH=8.0), 1 mM EDTA). To these resuspended cells, 36 µL of ssDNA mix (8 mg/mL Salmon Sperm DNA, 10 mM Tris (pH=8.0), 1 mM EDTA), 180 µL of β-

mercaptoethanol mix (5% β-mercaptoethanol, 95% triacetin), and 8 µg of plasmid library DNA were added, mixed via pipetting, and incubated for 30 mins. at room temperature. After incubation, 1800 µL of PEG mix (70% w/v PEG (3,350 MW)) was added and mixed via pipetting, and the mixture was incubated at room temperature for an additional 30 min. Cells were then heat shocked for 25 min at 37 °C, washed with 25 mL of sterile milliQ H_2O, and used to inoculate 50 mL of SD-leu media for screening experiments. Dilutions of the transformation (0.01% and 0.001%) were plated on solid SD-leu media to calculate transformation efficiency. Three biological replicates of each transformation were performed for each condition. Transformation efficiency for each replicate is presented in Supplementary Table S4.9. Details of the Cas9 library are provided in ref. [8]

Screening experiments were conducted in 50 mL of liquid media in a 250 mL baffled flask (220 rpm shaking, 30 °C). Cells first reached confluency after 2 days of growth (OD_{600} ~12), at which time 200 µL (which includes sufficient number of cells for approximately 500-fold library coverage) was used to inoculate 25 mL of fresh media. The cells were again subcultured upon reaching confluency at day 4 for the growth screen, and the experiment was halted after 6 days of growth. At each timepoint (*i.e.*, days 2, 4, and 6), 1 mL of culture was removed and treated with DNase I (New England Biolabs; 4 µL and 25µL of DNaseI buffer) for 1 h at 30 °C to remove any extracellular DNA. Cells were isolated by centrifugation at 4,500g and the resulting cell pellets were stored at -80 °C for future analysis.

4.5.7 Library isolation and sequencing

Growth screen samples were thawed and resuspended in 400 µL sterile, milliQ H_2O. Each cell suspension was split into two, 200 µL samples and plasmids from each sample were isolated using a Zymo Yeast Miniprep Kit (Zymo Research). Splitting into separate samples here was done to accommodate the capacity of the Yeast Miniprep Kit. The split samples from a single pellet were then pooled, and plasmid copy number was quantified using quantitative PCR with qPCR-GW-F and qPCR-GW-R and SsoAdvanced Universal SYBR Green Supermix (Biorad). Each pooled sample was confirmed to contain at least 10^7 plasmids.

To prepare samples for next generation sequencing, isolated plasmids were subjected to PCR using forward (ILU1-F, ILU2-F, ILU3-F, ILU4-F) and reverse primers (ILU(1-12)-R) containing all necessary barcodes and adapters for next generation sequencing using the Illumina platform (Supplementary Table S4.10). Schematics of the amplicons from the Cas9 and Cas12a experiments submitted for NGS are pictured in Supplementary Figure S4.8. At least 0.2 ng of plasmids (approximately $3x10^7$ plasmid molecules) were used as template, and PCR reactions were amplified for 16 cycles and not allowed to proceed to completion to avoid amplification bias. PCR product was purified using SPRI beads and tested on the bioanalyzer to ensure the correct length. Samples were pooled in equimolar amounts and submitted for sequencing on a NextSeq 500 at the UCR IIGB core facility.

4.5.8 Generating sgRNA read counts from raw reads

Next generation sequencing reads were processed using the Galaxy platform [45]. First read quality was assessed using FastQC v0.11.8. The reads were then demultiplexed using Cutadapt v1.16.6, trimmed using Trimmomatic v0.38, and mapped to each sgRNA using a combination of Bowtie 2 v2.4.2, and custom MATLAB scripts for counting bowtie alignments and naïve exact matching. Parameters used for each method are provided in Supplementary Table S4.11 and MATLAB scripts are provided as part of the GitHub link found below in the section "Data and software availability". Supplementary Table S4.12 provides further information correlating the NCBI SRA file names to the information needed for demultiplexing the readsets. Analysis of the CRISPR-Cas12a growth screens revealed that five sgRNAs were not present in the sequencing data. Pairwise comparison between normalized read abundances for biological replicates were done to verify consistency, see Supplementary Figure S4.2 and Supplementary Table S4.1.

4.5.9 Cutting Score analysis

The cutting score (CS) associated with each guide was determined by taking the log2 of the ratio of normalized read counts of the control condition to the normalized read counts of the treatment condition. The control condition was taken as the normalized read counts at the end of the growth screen in a strain without Cas12a or Cas9. The treatment condition included constitutively expressed Cas9 or Cas12a with disrupted KU70. Normalized counts were taken as the total number of reads for a given sgRNA divided by the total reads for the corresponding sample. If no reads were identified for a given sgRNA, a pseudo-count of one was added to the read count to facilitate subsequent calculations. In all cases,

normalized read counts for each biological replicate were averaged together to produce an average normalized read count and associated standard deviation for each sgRNA. All normalized read counts are provided in Supplementary Files 4.3 and 4.4.

4.5.10 Nucleosome Occupancy analysis

To account for genomic features, specifically nucleosome occupancy, we determined an average normalized occupancy score (ranging from 0 to 1) for every target locus using previously published MNase-Seq coverage data [46] (Supplementary File 4.5). Per base nucleosome occupancy scores were summed up for each sgRNA, averaged and normalized to a value between 0 and 1 by taking its ratio to the highest averaged value. This information was integrated into DeepGuide via a separate fully connected neural network, the first step of which was to convert the one-dimensional occupancy data into an 80-dimensional real vector using a fully connected layer with 80 neurons. Using element-wise multiplication, the output of this layer was combined with the output of the last fully connected layer of the CS-predicting CNN to generate CS predictions that account for guide sequence, genomic context, and nucleosome occupancy.

4.5.11 Validation of predicted sgRNA for Cas9 and Cas12a

Four genes with easily screenable phenotypes, including MEF1, CAN1, MGA1, and RAS2 were selected for the validation of predicted sgRNA CS values (Supplementary Figure S4.3). Gene sequences and the per base nucleosome occupancy of these genes were provided as input to the DeepGuide algorithm. As output DeepGuide predicted a CS value for each sgRNA of a given gene. sgRNAs were sorted from best to worst based on the

predicted CS value from sequence-only (for Cas12a) and sequence plus nucleosome occupancy (for Cas9). The top 5 and bottom 5 sgRNA from the list were tested for editing efficiency.

To screen for RAS2 and MGA1 gene disruption, cultures with CRISPR plasmids growing in SD-Leu were diluted and plated in triplicate on YPD to obtain greater than 50 colonies on each plate. After two days of growth at 30 °C, the number of smooth colonies were counted and expressed as a fraction of total colonies on the plate. For disruption of the CAN1 gene, cultures were similarly diluted and plated on YPD to obtain single colonies. Thirty colonies in triplicate were then randomly selected and streaked on SD-leu agar media supplemented with 50 mg L^{-1} of L-canavanine. Colonies that grew on SD with canavanine were identified as positive for CAN1 disruption. To screen for MFE1, cultures were similarly plated, and 30 colonies from each transformation were randomly selected and streaked on SD-Oleic acid and dotted on YPD. Growth on YPD but not on SD-Oleic acid indicated MFE1 disruption. Screening of MFE1 was done on agar plates containing SD media supplemented with oleic acid as the sole carbon source (SD oleic acid; 0.67% Difco yeast nitrogen base without amino acids, 0.079% CSM (Sunrise Science, San Diego, CA), 2% agar 0.4% (v/v) Tween 20, and 0.3% (v/v) oleic acid).

4.6 References

1. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337, 816–821 (2012).

2. Zetsche, B. et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell 163, 759–771 (2015).

3. Sadhu, M. J. et al. Highly parallel genome variant engineering with CRISPR–Cas9. Nature Genetics vol. 50 510–514 (2018).

4. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819–823 (2013).

5. Gilbert, L. A. et al. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell 159, 647–661 (2014).

6. Schwartz, C. M., Hussain, M. S., Blenner, M. & Wheeldon, I. Synthetic RNA Polymerase III Promoters Facilitate High-Efficiency CRISPR-Cas9-Mediated Genome Editing in Yarrowia lipolytica. ACS Synth. Biol. 5, 356–359 (2016).

7. Löbs, A.-K., Schwartz, C., Thorwall, S. & Wheeldon, I. Highly Multiplexed CRISPRi Repression of Respiratory Functions Enhances Mitochondrial Localized Ethyl Acetate Biosynthesis in Kluyveromyces marxianus. ACS Synth. Biol. 7, 2647–2655 (2018).

8. Schwartz, C. et al. Validating genome-wide CRISPR-Cas9 function improves screening in the oleaginous yeast Yarrowia lipolytica. Metab. Eng. 55, 102–110 (2019).

9. Liu, R., Chen, L., Jiang, Y., Zhou, Z. & Zou, G. Efficient genome editing in filamentous fungus Trichoderma reesei using the CRISPR/Cas9 system. Cell Discov 1, 15007 (2015).

10. Dalvie, N. C. *et al.* Host-Informed Expression of CRISPR Guide RNA for Genomic Engineering in Komagataella phaffii. *ACS Synthetic Biology* vol. 9 26–35 (2020).

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

12. Fuller, K. K., Chen, S., Loros, J. J. & Dunlap, J. C. Development of the CRISPR/Cas9 System for Targeted Gene Disruption in Aspergillus fumigatus. *Eukaryot. Cell* **14**, 1073–1080 (2015).

13. Cao, M., Gao, M., Ploessl, D., Song, C. & Shao, Z. CRISPR-Mediated Genome Editing and Gene Repression in Scheffersomyces stipitis. *Biotechnol. J.* **13**, e1700598 (2018).

14. Tran, V. G., Cao, M., Fatma, Z., Song, X. & Zhao, H. Development of a CRISPR/Cas9-Based Tool for Gene Deletion in Issatchenkia orientalis. *mSphere* vol. 4 (2019).

15. Moreb, E. A. & Lynch, M. D. Genome dependent Cas9/gRNA search time underlies sequence dependent gRNA activity. *Nat. Commun.* **12**, 5034 (2021).

16. Schwartz, C., Frogue, K., Misa, J. & Wheeldon, I. Host and Pathway Engineering for Enhanced Lycopene Biosynthesis in. *Front. Microbiol.* **8**, 2233 (2017).

17. Rodriguez, G. M. *et al.* Engineering xylose utilization in Yarrowia lipolytica by understanding its cryptic xylose pathway. *Biotechnol. Biofuels* **9**, 149 (2016).

18. Blazeck, J. *et al.* Harnessing Yarrowia lipolytica lipogenesis to create a platform for lipid and biofuel production. *Nat. Commun.* **5**, 3131 (2014).

19. Xue, Z. *et al.* Production of omega-3 eicosapentaenoic acid by metabolic engineering of Yarrowia lipolytica. *Nat. Biotechnol.* **31**, 734–740 (2013).

20. Lv, Y., Marsafari, M., Koffas, M., Zhou, J. & Xu, P. Optimizing Oleaginous Yeast Cell Factories for Flavonoids and Hydroxylated Flavonoids Biosynthesis. *ACS Synth. Biol.* **8**, 2514–2523 (2019).

21. Ledesma-Amaro, R., Dulermo, R., Niehus, X. & Nicaud, J.-M. Combining metabolic engineering and process optimization to improve production and secretion of fatty acids. *Metab. Eng.* **38**, 38–46 (2016).

22. Löbs, A.-K., Schwartz, C. & Wheeldon, I. Genome and metabolic engineering in non-conventional yeasts: Current advances and applications. *Synthetic and Systems Biotechnology* vol. 2 198–207 (2017).

23. Schwartz, C., Frogue, K., Ramesh, A., Misa, J. & Wheeldon, I. CRISPRi repression of nonhomologous end-joining for enhanced genome engineering via homologous recombination in Yarrowia lipolytica. *Biotechnol. Bioeng.* **114**, 2896–2906 (2017).

24. Chuai, G. *et al.* DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol.* **19**, 80 (2018).

25. Kim, H. K. *et al.* Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nat. Biotechnol.* **36**, 239–241 (2018).

26. Magnan, C. *et al.* Sequence Assembly of Yarrowia lipolytica Strain W29/CLIB89 Shows Transposable Element Diversity. *PLoS One* **11**, e0162363 (2016).

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

28. Labuhn, M. *et al.* Refined sgRNA efficacy prediction improves large- and small-scale CRISPR-Cas9 applications. *Nucleic Acids Res.* **46**, 1375–1385 (2018).

29. Xu, H. *et al.* Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **25**, 1147–1157 (2015).

30. Chari, R., Yeo, N. C., Chavez, A. & Church, G. M. sgRNA Scorer 2.0: A Species-Independent Model To Predict CRISPR/Cas9 Activity. *ACS Synth. Biol.* **6**, 902–904 (2017).

31. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

32. Peng, H., Zheng, Y., Blumenstein, M., Tao, D. & Li, J. CRISPR/Cas9 cleavage efficiency regression through boosting algorithms and Markov sequence profiling. *Bioinformatics* **34**, 3069–3077 (2018).

33. Xiang, X. *et al.* Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. *Nat. Commun.* **12**, 3238 (2021).

34. Horlbeck, M. A. *et al.* Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *Elife* **5**, (2016).

35. Yarrington, R. M., Verma, S., Schwartz, S., Trautman, J. K. & Carroll, D. Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 9351–9358 (2018).

36. Strohkendl, I. *et al.* Inhibition of CRISPR-Cas12a DNA targeting by nucleosomes and chromatin. *Sci Adv* **7**, (2021).

37. Verkuijl, S. A. & Rots, M. G. The influence of eukaryotic chromatin state on CRISPR-Cas9 editing efficiencies. *Curr. Opin. Biotechnol.* **55**, 68–73 (2019).

38. Guo, J. *et al.* Improved sgRNA design in bacteria via genome-wide activity profiling. *Nucleic Acids Res.* **46**, 7052–7069 (2018).

39. Wang, D. *et al.* Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* **10**, 4284 (2019).

40. Hasan, M. A. & Lonardi, S. DeeplyEssential: a deep neural network for predicting essential genes in microbes. *BMC Bioinformatics* **21**, 367 (2020).

41. Beder, T. *et al.* Identifying essential genes across eukaryotes by machine learning. *NAR Genom Bioinform* **3**, lqab110 (2021).

42. Ramesh, A. & Wheeldon, I. Guide RNA Design for Genome-Wide CRISPR Screens in Yarrowia lipolytica. *Methods Mol. Biol.* **2307**, 123–137 (2021).

43. Schwartz, C., Shabbir-Hussain, M., Frogue, K., Blenner, M. & Wheeldon, I. Standardized Markerless Gene Integration for Pathway Engineering in Yarrowia lipolytica. *ACS Synth. Biol.* **6**, 402–409 (2017).

44. Ramesh, A., Ong, T., Garcia, J. A., Adams, J. & Wheeldon, I. Guide RNA Engineering Enables Dual Purpose CRISPR-Cpf1 for Simultaneous Gene Editing and Gene Regulation in. *ACS Synth. Biol.* **9**, 967–971 (2020).

45. Jalili, V. *et al.* Corrigendum: The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.* **48**, 8205–8207 (2020).

46. Tsankov, A. M., Thompson, D. A., Socha, A., Regev, A. & Rando, O. J. The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol.* **8**, e1000414 (2010).

4.7 Data availability

The sgRNA sequencing data generated in this study have been deposited in the NCBI SRA database under accession code PRJNA766088 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA766088]. The sgRNA activity data (cutting scores) generated in this study are provided in the Supplementary Information/Source Data file.

4.8 Code availability

Source code for DeepGuide can be found at https://github.com/dDipankar/DeepGuide. Our GitHub page includes instructions for installation, usage examples. Custom MATLAB

scripts that were used for the design of the Cas12a CRISPR library, and processing of Illumina reads to generate sgRNA abundance can also be found in the GitHub page. The Github repository has been archived to Zenodo to provide a permanent reference to the version of code used in this study [https://doi.org/10.5281/zenodo.5889577]. Generating sgRNA predictions for *Y. lipolytica* using DeepGuide does not require any specialized hardware and it can be carried out on a laptop with Conda installed.

4.9 Author contributions statement

All authors conceived the idea and wrote the manuscript. AR, CS, and IW planned and analyzed the genome-wide CRISPR screens. AR conducted the CRISPR-Cas12a and guide-activity validation experiments. CS conducted the CRISPR-Cas9 screens. DB and SL planned the computational prediction of guide activity. DB designed and optimized the architecture of DeepGuide, and collected data with DeepGuide and all other sgRNA prediction tools.

4.10 Competing interests statement

The authors declare no competing interests.

4.11 Acknowledgements

This work was supported by DOE DE-SC0019093 (IW and SL), DOE Joint Genome Institute grant CSP-503076 (IW) and NSF 1706545 (IW).

4.12 Supplementary Information

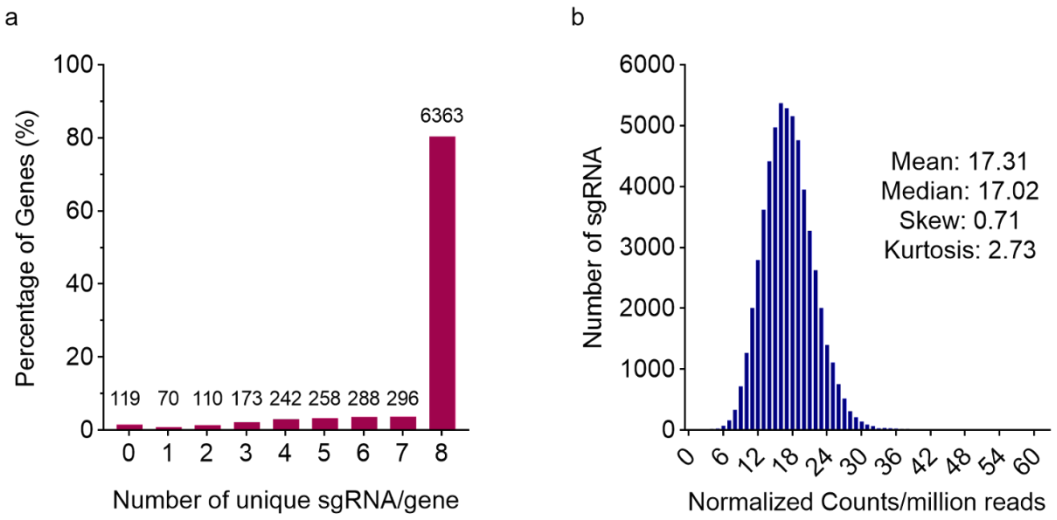

Figure S4.1. **Design and validation of Cas12a and Cas9 sgRNA library for *Y. lipolytica* PO1f.** (a) An 8-fold redundant sgRNA library was designed to target 7,919 protein coding genes in the *Y. lipolytica* CLIB89 strain, the parent strain of PO1f. Coding sequences were confirmed to be present in the PO1f genome sequence. Over 80% of the genes had 8 sgRNAs and over 91% of the genes had at least 5 sgRNAs. (b) A library consisting of 58,421 sgRNAs was synthesized by Agilent, cloned in-house and characterized by next generation sequencing. The library exhibited a tight normal distribution with nearly equal mean and median signifying minimal skew. The average representation of sgRNAs was ~100-fold (at 5.84 million reads which is 100 times the library size, we can calculate the mean representation of sgRNAs to be 5.84*17.31 = 101.09). Note: The Cas9 library design was previously reported in ref. 1 (see Figure S1). Additional details of this library are also provided in the materials and methods section of this manuscript.

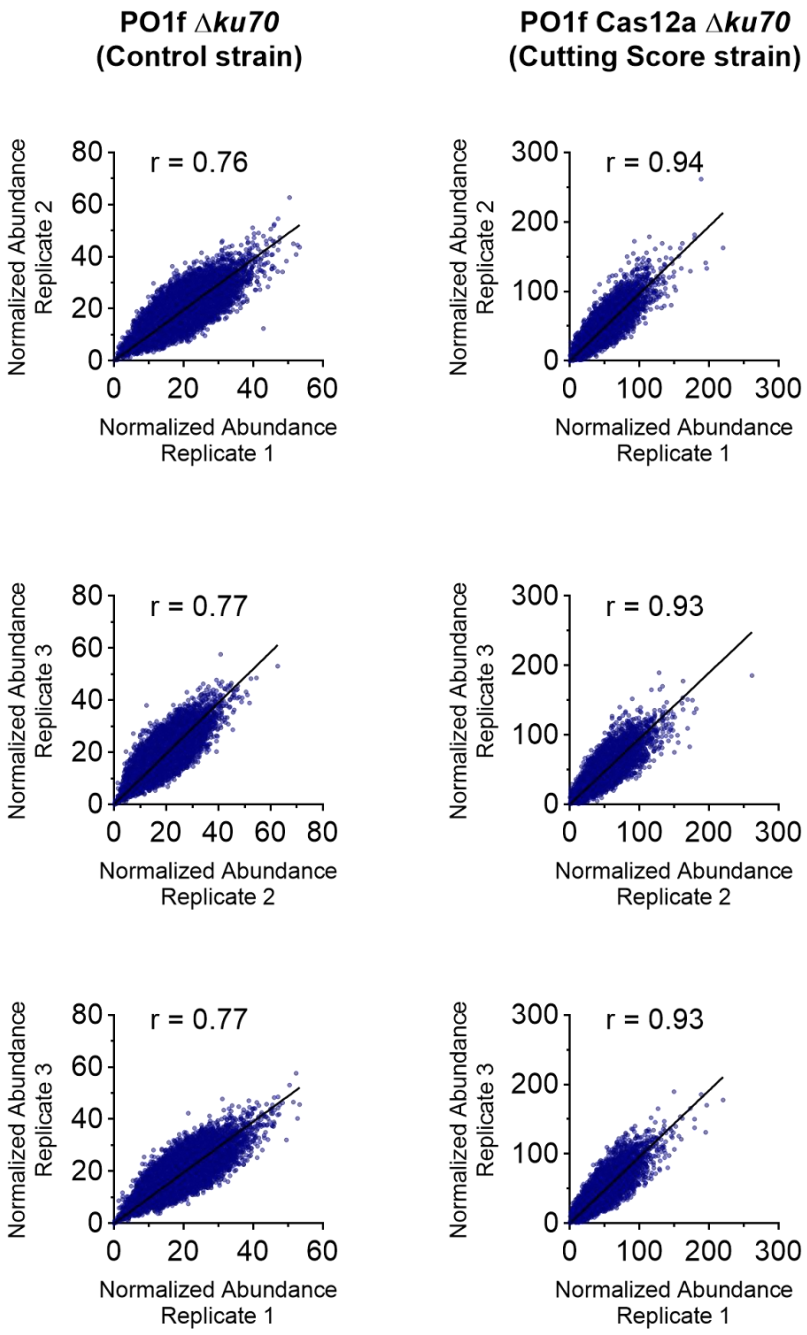

Figure S4.2. Replicate correlation graphs at Day 4 of the growth screen for Cas12a experiments. The column on the left shows pairwise correlations for the control strain while the column on the right shows the same for sample strain.

Table S4.1. Replicate correlations for the genome-wide growth screens in *Y. lipolytica* with the Cas9 and Cas12a endonucleases. Cas9 data was previously reported in ref. [1] Note: Work conducted in ref. 1 uses PO1f with functional KU70 as the control strain.

Strain	Time point	Comparison	Pearson
PO1f *ku70*	Day 2	1 v. 2	0.765
		1 v. 3	0.775
		2 v. 3	0.738
	Day 4	1 v. 2	0.756
		1 v. 3	0.772
		2 v. 3	0.762
	Day 6	1 v. 2	0.797
		1 v. 3	0.768
		2 v. 3	0.799
PO1f Cas12a *ku70*	Day 2	1 v. 2	0.902
		1 v. 3	0.925
		2 v. 3	0.892
	Day 4	1 v. 2	0.936
		1 v. 3	0.933
		2 v. 3	0.927
	Day 6	1 v. 2	0.918
		1 v. 3	0.915
		2 v. 3	0.905

Strain	Time point	Comparison	Pearson
PO1f	Day 2	1 v. 2	0.988
		1 v. 3	0.982
		2 v. 3	0.980
	Day 4	1 v. 2	0.829
		1 v. 3	0.827
		2 v. 3	0.858
	Day 6	1 v. 2	0.818
		1 v. 3	0.829
		2 v. 3	0.855
PO1f Cas9 *ku70*	Day 2	1 v. 2	0.972
		1 v. 3	0.976
		2 v. 3	0.972
	Day 4	1 v. 2	0.886
		1 v. 3	0.891
		2 v. 3	0.973
	Day 6	1 v. 2	0.877
		1 v. 3	0.875
		2 v. 3	0.968

Table S4.2. The twelve layers in the convolutional auto-encoder (first network in DeepGuide); the autoencoder is composed by an encoder (layers 1-6) and a decoder (layers 7-12).

CAE (1st network)	Layer #	Layer type
Encoder	1	Convolution
	2	Batch Normalization
	3	Max Pooling
	4	Convolution
	5	Batch Normalization
	6	Average Pooling
Decoder	7	Up Sampling
	8	Batch Normalization
	9	Convolution
	10	Up Sampling
	11	Batch Normalization
	12	Convolution

Table S4.3. The eleven layers in the second network in DeepGuide, composed of an encoder (layers 1-6) and a fully connected network (layers 7-11).

2nd network	Layer #	Layer type
Encoder	1	Convolution
	2	Batch Normalization
	3	Max Pooling
	4	Convolution
	5	Batch Normalization
	6	Average Pooling
Fully connected network	7	Flatten
	8	Fully connected
	9	Fully connected
	10	Fully connected
	11	Multiplication

Table S4.4. Ablation analysis on Cas12a dataset. Green row (row 1) show the performance of the encoder (followed by a flatten layer) using random weights (no pre-training or backpropagation); purple row (row 2) show the performance of the encoder (followed by a flatten layer) using random weights and then performing back-propagation only on the flatten layer; blue rows (3-7) show the performance after pre-training the encoder and then running back-propagation only layers downstream of the encoder; pink rows (8-12) show the performance after pre-training and then running back-propagation on the whole network (including the encoder); correlation coefficients in bold corresponds to the best performance; fc = fully connected layer; pool = pooling layer; flatten = flatten layer; mult = multiplication layer (see Table S5 for the list of layers)

Cas12a	Layers	Spearman	Pearson
No pre-training (random weights), **no** back-propagation	encoder⇨flatten$_{7}$	0.060	0.070
No pre-training (random weights), followed by back-propagation **only** on the flatten layer	encoder⇨flatten$_{7}$	0.451	0.455
Pre-training of the encoder followed by back-propagation **only** the layers downstream of the encoder (flatten$_{7}$⇨…)	encoder⇨flatten$_{7}$	0.521	0.532
	encoder⇨flatten$_{7}$⇨fc$_{8}$	**0.527**	**0.534**
	encoder⇨flatten$_{7}$⇨fc$_{8}$⇨fc$_{9}$	0.505	0.517
	encoder⇨flatten$_{7}$⇨fc$_{8}$⇨fc$_{9}$⇨fc$_{10}$	0.501	0.514
	encoder⇨flatten$_{7}$⇨fc$_{8}$⇨fc$_{9}$⇨fc$_{10}$⇨mult$_{11}$	0.501	0.514
Pre-training of the encoder followed by back-propagation on the entire network	encoder⇨flatten$_{7}$	0.637	0.641
	encoder⇨flatten$_{7}$⇨fc$_{8}$	0.649	0.658
	encoder⇨flatten$_{7}$⇨fc$_{8}$⇨fc$_{9}$	**0.653**	**0.660**
	encoder⇨flatten$_{7}$⇨fc$_{8}$⇨fc$_{9}$⇨fc$_{10}$	0.653	0.660
	encoder⇨flatten$_{7}$⇨fc$_{8}$⇨fc$_{9}$⇨fc$_{10}$⇨mult$_{11}$	0.653	0.660

Table S4.5. Ablation analysis on Cas9 dataset. Green row (row 1) show the performance of the encoder (followed by a flatten layer) using random weights (no pre-training or backpropagation); purple row (row 2) show the performance of the encoder (followed by a flatten layer) using random weights and then performing back-propagation only on the flatten layer; blue rows (3-7) show the performance after pre-training the encoder and then running back-propagation only layers downstream of the encoder; pink rows (8-12) show the performance after pre-training and then running back-propagation on the whole network (including the encoder); correlation coefficients in bold corresponds to the best performance; fc = fully connected layer; pool = pooling layer; flatten = flatten layer; mult = multiplication layer (see Table S5 for the list of layers)

Cas9	Layers	Spearman r	Pearson r
No pre-training (random weights), **no** back-propagation	encoder⇨flatten$_7$	0.004	0.003
No pre-training (random weights), followed by back-propagation **only** on the flatten layer	encoder⇨flatten$_7$	0.291	0.312
Pre-training of the encoder followed by back-propagation **only** the layers downstream of the encoder (flatten$_7$⇨…)	encoder⇨flatten$_7$	0.316	0.353
	encoder⇨flatten$_7$ ⇨fc$_8$	0.273	0.310
	encoder⇨flatten$_7$ ⇨fc$_8$⇨fc$_9$	0.261	0.291
	encoder⇨flatten$_7$ ⇨fc$_8$⇨fc$_9$⇨fc$_{10}$	0.269	0.305
	encoder⇨flatten$_7$ ⇨fc$_8$⇨fc$_9$⇨fc$_{10}$⇨mult$_{11}$	**0.345**	**0.388**
Pre-training of the encoder followed by back-propagation on the entire network	encoder⇨flatten$_7$	0.347	0.409
	encoder⇨flatten$_7$ ⇨fc$_8$	0.364	0.424
	encoder⇨flatten$_7$ ⇨fc$_8$⇨fc$_9$	0.357	0.414
	encoder⇨flatten$_7$ ⇨fc$_8$⇨fc$_9$⇨fc$_{10}$	0.357	0.414
	encoder⇨flatten$_7$ ⇨fc$_8$⇨fc$_9$⇨fc$_{10}$⇨mult$_{11}$	**0.431**	**0.501**

Gene Name	Function	Observed phenotype of null
MGA1	Heat shock factor & pseudohyphal growth	Smooth colonies
RAS2	GTP-binding protein, regulates filamentous growth	Smooth colonies
CAN1	Arginine permease	Canavanine resistance
MFE1	β-oxidation of long chain fatty acids	Oleic acid metabolism muted

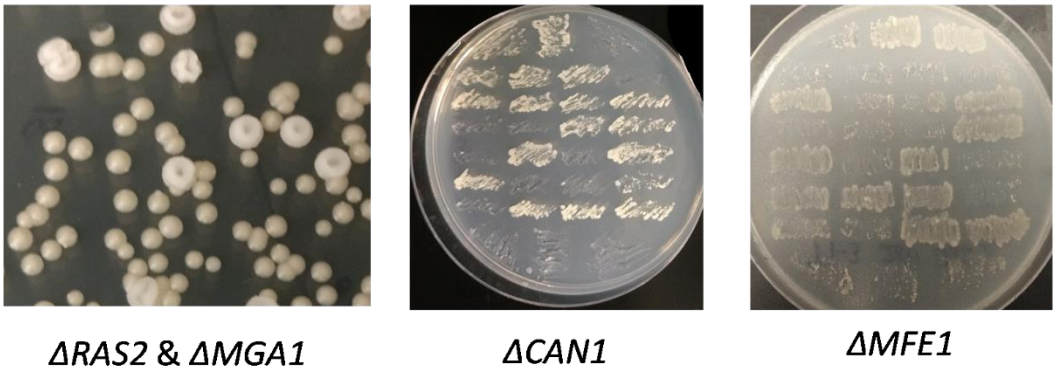

Figure S4.3. Genes selected for experimental validation of DeepGuide and the observed phenotype of the null mutants. MGA1 and RAS2 are implicated in the pseudohyphal and filamentous growth, and their null mutants show smooth colonies as shown in the picture on the left. CAN1 disruption confers resistance to L-Canavanine which is a toxic analog of Arginine. This leads to growth on plates supplemented with canavanine, as shown in the middle picture. MFE1 disruption renders *Y. lipolytica* unable to utilize oleic acid as a carbon source, and null mutants do not grow on plates with oleic acid as the sole carbon source as shown in the right-most picture.

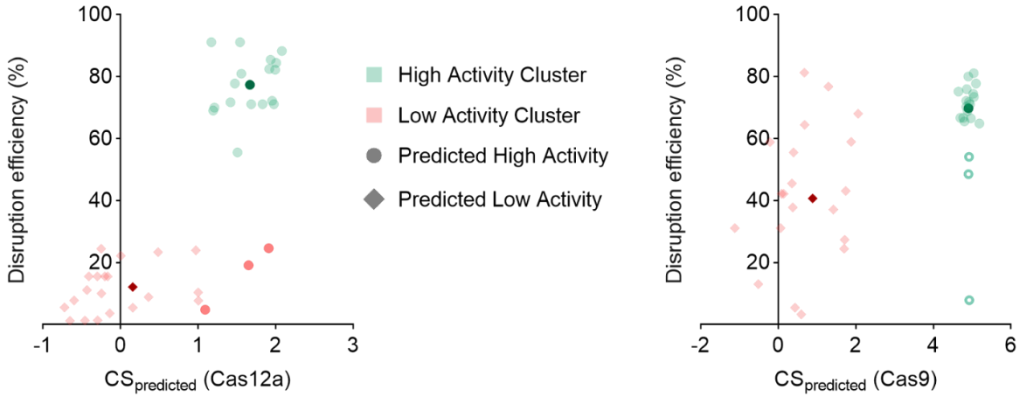

Figure S4.4. Clustering of high and poor activity guides used to validate DeepGuide. Predicted CS values and experimental disruption efficiencies for both the Cas12a and Cas9 were plotted on an XY scatter plot and a gaussian mixture model was used to cluster the sgRNA into two clusters (high and low activity). The high activity clusters are indicated in green, while the low activity clusters are indicated in red. Dark green and red points correspond to cluster centroids. Data point shape indicates whether the guide was predicted to be of high or low activity (circles are high activity, diamonds are low activity). Three predicted high activity guides cluster with low activity guides for Cas12a. For Cas9, three guides in the high activity cluster have a significantly higher euclidean distance from the cluster centroid and appear to be outliers (marked with empty circles).

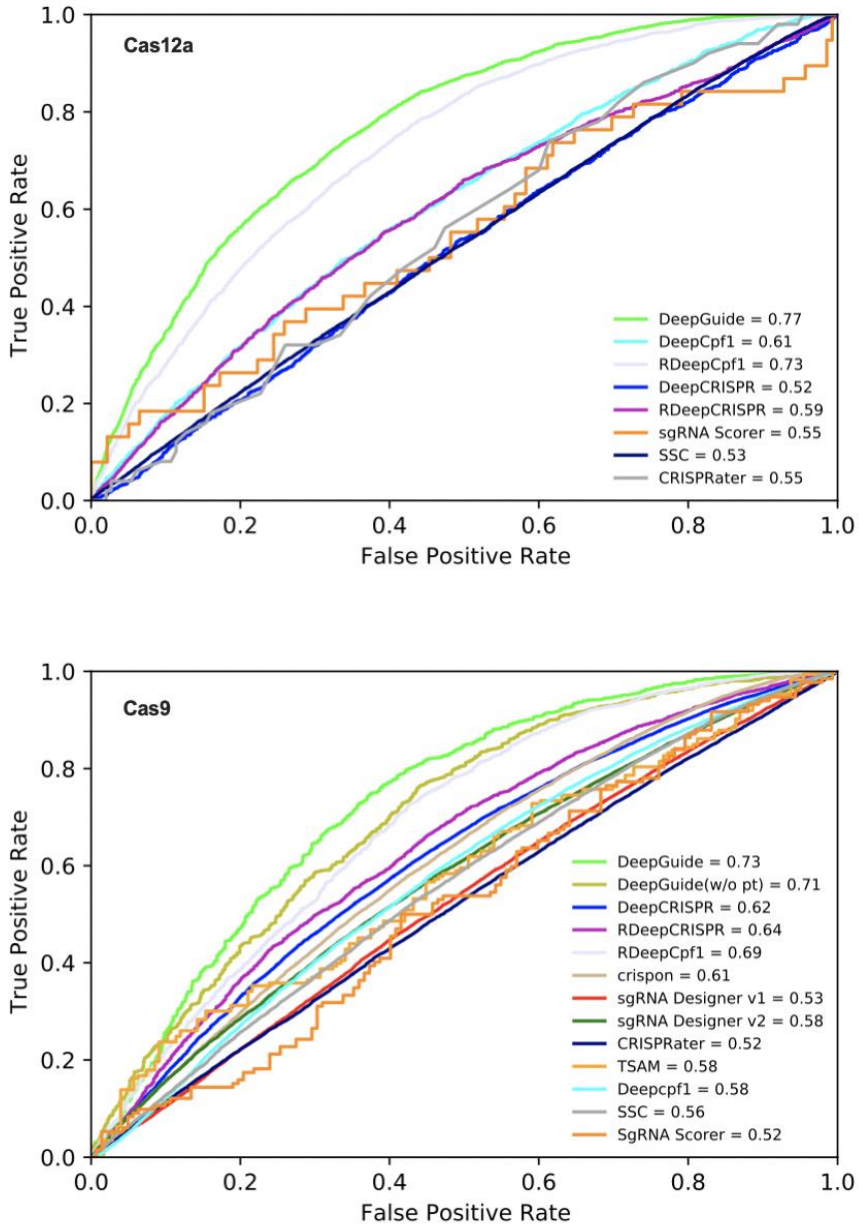

Figure S4.5. ROC plots and AUROC values for DeepGuide, DeepCpf1 (original and retrained), DeepCRISPR (original and retrained), sgRNA Scorer, SSC, and CRISPRater for the prediction of sgRNA activity on the Cas12a dataset (left) and the Cas9 dataset (right). DeepGuide had higher AUROC values than all other guide activity prediction algorithms. Guides with CS > 1.67 for Cas12a and CS > 4.91 for Cas9 were classified as active, and guides with a CS value below this threshold were classified as inactive. DeepGuide (w/o pt) indicates that no pre-training was carried out.

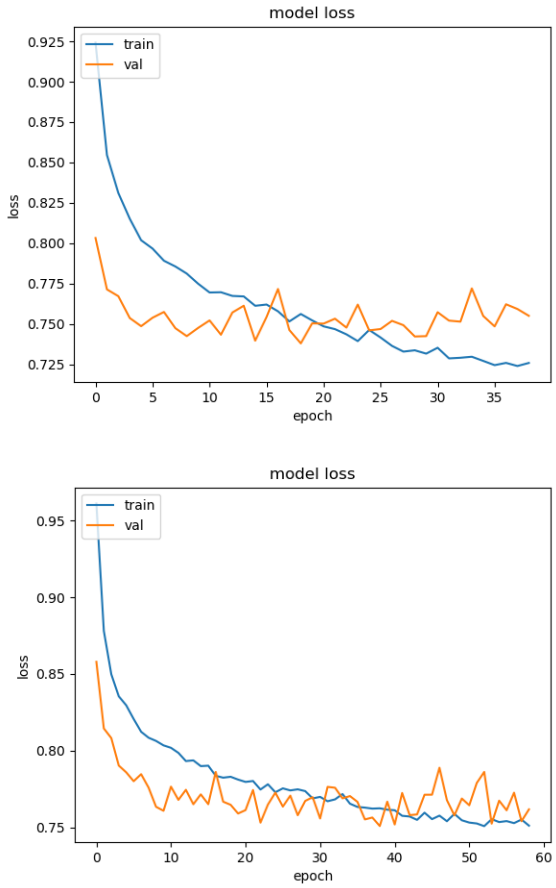

Figure S4.6. Training and validation loss for DeepGuide without pre-training (left) and with pre-training (right) as a function of the number of training epochs. These curves show that pre-training improves the architecture's generalization.

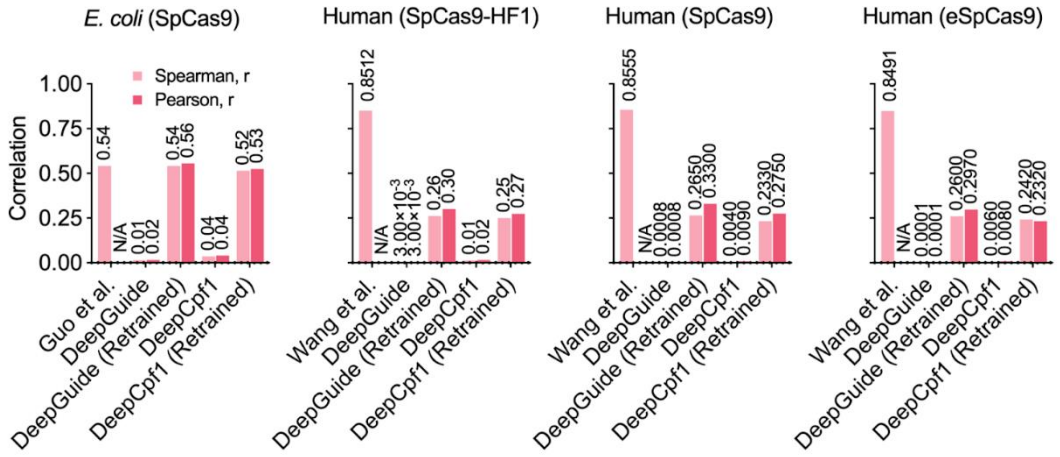

Figure S4.7. Evaluation of DeepGuide's ability to predict guide activity in other species. DeepGuide was tested on four non-Yarrowia datasets, including a CRISPR-Cas9 activity profile in *E. coli* [2] and three CRISPR-Cas9 datasets in mammalian cell lines [3]. These datasets were selected from the 44 publicly available sets listed in ref. 4, because they were the only ones having a size comparable to our *Y. lipolytica* datasets (*i.e.*, they contained at least 30,000 data points; see Figure 4 of main text, DeepGuide requires at least this many data points for high accuracy predictions). DeepGuide, before and after retraining, was compared to DeepCpf1 [5] (also before and after retraining) on all four datasets, as well as to the method originally developed for the respective datasets. DeepCpf1 was chosen because of its strong performance on our *Y. lipolytica* datasets. The data show that (i) retraining is necessary for DeepGuide and DeepCpf1 to achieve a reasonable predictive performance, (ii) when retrained, DeepGuide achieves a slightly higher predictive performance than DeepCpf1. We note that the Spearman coefficient reported on the *E. coli* dataset using the method proposed in ref. 2, which is based on gradient boosting regression trees, was 0.542. This matches the performance of DeepGuide, showing that our method is able to capture CRISPR-Cas9 activity in *E. coli*. DeepGuide was not able to capture guide activity measured in mammalian cell lines, thus demonstrating the importance of architecture optimization for broad cross-species prediction abilities.

Table S4.6. Yeast strains used in this study.

Yeast strain genotype	Phenotype
PO1f (MatA, *leu2-270*, *ura3-302*, *xpr2-322*, *axp-2*)	Wild type strain
PO1f *Δku70*	PO1f with disrupted KU70, which facilitates the non-homologous end joining DNA repair pathway
PO1f UAS1B8-TEF(136)-Cas9 -CycT::A08	PO1f expressing *Y. lipolytica* codon optimized Cas9 gene at the A08 locus
PO1f UAS1B8-TEF(136)-LbCas12a -CycT::A08	PO1f expressing *Y. lipolytica* codon optimized LbCas12a gene at the A08 locus
PO1f *Δku70* UAS1B8-TEF(136)-Cas9 -CycT::A08	KU70 disrupted in Cas9 integrated PO1f strain
PO1f *Δku70* UAS1B8-TEF(136)-LbCas12a -CycT::A08	KU70 disrupted in LbCas12a integrated PO1f strain

Table S4.7. Plasmids used for genome wide CRISPR screens.

Plasmid name	Reference	Function
pCpf1_yl	[6]	Plasmid for CRISPR-LbCas12a based gene editing in *Y. lipolytica*
pCRISPRyl	[7]	Plasmid for CRISPR-Cas9 based gene editing in *Y. lipolytica*
pLbCas12ayl	This study	Plasmid for CRISPR-LbCas12a based gene editing in *Y. lipolytic*a. sgRNA is flanked on either end by the direct repeat, to allow sgRNAs to end in T residues without being construed as part of the PolyT terminator
pHR_A08_hrGFP (Addgene #84615)	This study	Plasmid containing homology arms for integration of hrGFP into the A08 locus
pHR_A08_LbCas12a	This study	Plasmid containing homology arms for integration of LbCas12a into the A08 locus
pHR_A08_Cas9	[1]	Plasmid containing homology arms for integration of Cas9 into the A08 locus
pLbCas12ayl-GW	This study	Vector containing sgRNA expression cassette for cloning Cas12a sgRNA library. (Does not contain Cas12a expression cassette)
pCas9yl-GW	[1]	Vector containing sgRNA expression cassette for cloning Cas9 sgRNA library. (Does not contain Cas9 expression cassette)
pCRISPRyl_KU70	This study	CRISPR plasmid for the disruption of KU70

Table S4.8. Sequences of primers used in this study.

Primer name	Primer Sequence
ExtraDR-F	CGGCGCAAATTTCTACTAAGTGTAGACTAGTAATTTCTACTAAGTGTAGATTTTTTACGTCTAAGAAACCATTATT
ExtraDR-R	AATAATGGTTTCTTAGACGTAAAAAAATCTACACTTAGTAGAAATTACTAGTCTACACTTAGTAGAAATTTGCGCCG
Cpf1-Int-F	TGCCTGGAGCCGAGTACGGCATTGATTACTAGTCCGGGTTCGAAGGTACCAAG
Cpf1-Int-R	TTAGGCTGGGTCTCGAGAGCAAAGAAGCCTAGGGCAAATTAAAGCCTTCGAGCG
BRIDGE-F	CTAAATTTGATGAAAGGGGGATCCCCCGGGTGGCGTAATCATGGTCATAGCTGTTTCCTG
BRIDGE-R	CAGGAAACAGCTATGACCATGATTACGCCACCCGGGGGATCCCCCTTTCATCAAATTTAG
A08-Seq-F	AGCCGAGTACGGCATTGAT
A08-Seq-R	TCAATGTAGCCTCCTCCAACC
Tef_Seq-F	GTTGGGACTTTAGCCAAG
Lb1-R	CTTCTGCTTGGTCTTCTGGTTG
Lb2-F	AACCTGTACAACCAGAAGACCAAG
Lb3-F	AAGGAGACCAACCGAGACGAG
Lb4-F	AACCTGCACACCATGTACTTCAAG
Lb5-F	CCAGATCACCAACAAGTTCGAGTC
M13-F	GTAAAACGACGGCCAGT
InversePCR-F	TTTTTTTACGTCTAAGAAACCATTATTATCATGACATTAACCT
InversePCR-R	TGCGCCGACCCGGAATCGAACCGGGGGCCC
OLS-F	GTTTAGTGGTAAAATCCATCGTTGCCATCG
OLS-R	GATACGCCTATTTTTATAGGTTAATGTCATG
qPCR-GW-F	TTATGAACTGAAAGTTGATGGC
qPCR-GW-R	TCACACAGGAAACAGCTATG
Cas9-RAS2-1	TTCGATTCCGGGTCGGCGCACGCGGTCACTCCCCGCTCGTGTTTTAGAGCTAGAAATAGC
Cas9-RAS2-2	TTCGATTCCGGGTCGGCGCACTCCACCAGTGGAGCCAACCGTTTTAGAGCTAGAAATAGC
Cas9-RAS2-3	TTCGATTCCGGGTCGGCGCAACCTCCTGCAGCACCTCCAAGTTTTAGAGCTAGAAATAGC
Cas9-RAS2-4	TTCGATTCCGGGTCGGCGCAGACTCTCAATGCTCCACCAGGTTTTAGAGCTAGAAATAGC
Cas9-RAS2-5	TTCGATTCCGGGTCGGCGCAGATGTCGTAAACCAGAAGATGTTTTAGAGCTAGAAATAGC

Cas9-RAS2-6	TTCGATTCCGGGTCGGCGCAAATCTAGGGCCTCCAAAGACGTTTTAGAGCTAGAAATAGC
Cas9-RAS2-7	TTCGATTCCGGGTCGGCGCATCCCGTTCCTGTGGTTAGTAGTTTTAGAGCTAGAAATAGC
Cas9-RAS2-8	TTCGATTCCGGGTCGGCGCATGTTGGAGTCGACCTGGAAGGTTTTAGAGCTAGAAATAGC
Cas9-RAS2-9	TTCGATTCCGGGTCGGCGCAAAGCTGTGGGTGCACTGGTCGTTTTAGAGCTAGAAATAGC
Cas9-RAS2-10	TTCGATTCCGGGTCGGCGCAGGAACCAGAGGACTAAGCTGGTTTTAGAGCTAGAAATAGC
Cas9-MGA1-1	TTCGATTCCGGGTCGGCGCACTGTTGCGCGGCCTGGGTCGGTTTTAGAGCTAGAAATAGC
Cas9-MGA1-2	TTCGATTCCGGGTCGGCGCAACTGGCCAAGGAGCCTGCTGGTTTTAGAGCTAGAAATAGC
Cas9-MGA1-3	TTCGATTCCGGGTCGGCGCATTGCGGCAGAGGCATGGTTTGTTTTAGAGCTAGAAATAGC
Cas9-MGA1-4	TTCGATTCCGGGTCGGCGCACAGAGGCATGGTTTCGGCGCGTTTTAGAGCTAGAAATAGC
Cas9-MGA1-5	TTCGATTCCGGGTCGGCGCAGCCCGGCGAGGAGTTCTCCAGTTTTAGAGCTAGAAATAGC
Cas9-MGA1-6	TTCGATTCCGGGTCGGCGCAAAGACGGAGTTTGTGGGTGGGTTTTAGAGCTAGAAATAGC
Cas9-MGA1-7	TTCGATTCCGGGTCGGCGCAAGAGAGACAGTGTGCCCTTGGTTTTAGAGCTAGAAATAGC
Cas9-MGA1-8	TTCGATTCCGGGTCGGCGCAGTAGGGGGCGCCTGTCCGTCGTTTTAGAGCTAGAAATAGC
Cas9-MGA1-9	TTCGATTCCGGGTCGGCGCAGAGTGTGGTGGCGGAGTAGAGTTTTAGAGCTAGAAATAGC
Cas9-MGA1-10	TTCGATTCCGGGTCGGCGCATGCGCGGCCTGGGTCGTGGGGTTTTAGAGCTAGAAATAGC
Cas9-CAN1-1	TTCGATTCCGGGTCGGCGCATCAAACGATTACCCACCCTCGTTTTAGAGCTAGAAATAGC
Cas9-CAN1-2	TTCGATTCCGGGTCGGCGCATTACCCACCCTCCGGGACTGGTTTTAGAGCTAGAAATAGC
Cas9-CAN1-3	TTCGATTCCGGGTCGGCGCACCACATCCACATCAACCACAGTTTTAGAGCTAGAAATAGC
Cas9-CAN1-4	TTCGATTCCGGGTCGGCGCACATCAACCACACGGCCCACTGTTTTAGAGCTAGAAATAGC
Cas9-CAN1-5	TTCGATTCCGGGTCGGCGCACACCAGTGGCCACGACCTGGGTTTTAGAGCTAGAAATAGC
Cas9-CAN1-6	TTCGATTCCGGGTCGGCGCAAGTGGGCCGTGTGGTTGATGGTTTTAGAGCTAGAAATAGC
Cas9-CAN1-7	TTCGATTCCGGGTCGGCGCACCGTGTGGTTGATGTGGATGGTTTTAGAGCTAGAAATAGC
Cas9-CAN1-8	TTCGATTCCGGGTCGGCGCAGTGGATGTGGGCCTCAGTCCGTTTTAGAGCTAGAAATAGC

Cas9-CAN1-9	TTCGATTCCGGGTCGGCGCAGATGTGGGCCTCAGTCCCGGGTTTTAGAGCTAGAAATAGC
Cas9-CAN1-10	TTCGATTCCGGGTCGGCGCATGGGCCTCAGTCCCGGAGGGGTTTTAGAGCTAGAAATAGC
Cas9-MFE1-1	TTCGATTCCGGGTCGGCGCATGGTGAGACCCTGAAGGTTGGTTTTAGAGCTAGAAATAGC
Cas9-MFE1-2	TTCGATTCCGGGTCGGCGCAGGTGTTATCCCTTACATGGGGTTTTAGAGCTAGAAATAGC
Cas9-MFE1-3	TTCGATTCCGGGTCGGCGCACGTACTTCTGCTTAAGGAAGGTTTTAGAGCTAGAAATAGC
Cas9-MFE1-4	TTCGATTCCGGGTCGGCGCAGACAAGATCCCAGTCCTTGTGTTTTAGAGCTAGAAATAGC
Cas9-MFE1-5	TTCGATTCCGGGTCGGCGCAATACTTGAGCTCATTAGCCTGTTTTAGAGCTAGAAATAGC
Cas9-MFE1-6	TTCGATTCCGGGTCGGCGCACTGCTTTCGGAAGTAAGGCCGTTTTAGAGCTAGAAATAGC
Cas9-MFE1-7	TTCGATTCCGGGTCGGCGCAAAAGCAGGGTCGATGTGAAGGTTTTAGAGCTAGAAATAGC
Cas9-MFE1-8	TTCGATTCCGGGTCGGCGCAGTCGATGAAATTAAGGCCCTGTTTTAGAGCTAGAAATAGC
Cas9-MFE1-9	TTCGATTCCGGGTCGGCGCAGTTGTTGTCAACGATCTTGGGTTTTAGAGCTAGAAATAGC
Cas9-MFE1-10	TTCGATTCCGGGTCGGCGCACTTGGATCGGACAGACTCGAGTTTTAGAGCTAGAAATAGC
Cas12a-RAS2-1	TTTCTACTAAGTGTAGATGAGGCCCTAGATTACTTCAACGACAAATTTCTACTAAGTGTA
Cas12a-RAS2-2	TTTCTACTAAGTGTAGATGACCACCTAACGACGCGAAAAAACAAATTTCTACTAAGTGTA
Cas12a-RAS2-3	TTTCTACTAAGTGTAGATCGACATCACAGCCCCCCAGTCTTTGAATTTCTACTAAGTGTA
Cas12a-RAS2-4	TTTCTACTAAGTGTAGATGGCACCCGCACACCGGCCCCAGCTTAATTTCTACTAAGTGTA
Cas12a-RAS2-5	TTTCTACTAAGTGTAGATCATGAATCCGCATCCATGCTCGCGCAATTTCTACTAAGTGTA
Cas12a-RAS2-6	TTTCTACTAAGTGTAGATCATTGTCATTCTTGGAGAGGGAGGTAATTTCTACTAAGTGTA
Cas12a-RAS2-7	TTTCTACTAAGTGTAGATCGTCGCGACTGGGTGTGTCTGATCGAATTTCTACTAAGTGTA
Cas12a-RAS2-8	TTTCTACTAAGTGTAGATGCGTCGTTAGGTGGTCCAAAACGAGAATTTCTACTAAGTGTA
Cas12a-RAS2-9	TTTCTACTAAGTGTAGATCTGAAGTTTCCATGAATCCGCATCCAATTTCTACTAAGTGTA
Cas12a-RAS2-10	TTTCTACTAAGTGTAGATCGCGACTTTGCGCACTATAGATGAGAATTTCTACTAAGTGTA
Cas12a-MGA1-1	TTTCTACTAAGTGTAGATTGGGTGGTGGATTCGCTGAAGCGCTAATTTCTACTAAGTGTA

Cas12a-MGA1-2	TTTCTACTAAGTGTAGATATGGTCTGCGTCCAACGACTCGTTCAATTTCTACTAAGTGTA
Cas12a-MGA1-3	TTTCTACTAAGTGTAGATGGCGGCATGTGCTCGACCCGTTCTTAATTTCTACTAAGTGTA
Cas12a-MGA1-4	TTTCTACTAAGTGTAGATTGCGCCAGCTCAACATGTACGGCTTAATTTCTACTAAGTGTA
Cas12a-MGA1-5	TTTCTACTAAGTGTAGATGGTGGCCCATGGCGTGTGCCACCCGAATTTCTACTAAGTGTA
Cas12a-MGA1-6	TTTCTACTAAGTGTAGATTCAACAATCTGCAGCAGCGTCTGCAAATTTCTACTAAGTGTA
Cas12a-MGA1-7	TTTCTACTAAGTGTAGATTTGAACCCAGAAGGGGGCGACAAGAAATTTCTACTAAGTGTA
Cas12a-MGA1-8	TTTCTACTAAGTGTAGATGAGTGGTGCCGGGCTTCTTGTTATCTTTTTTACGTCTAAGAA
Cas12a-MGA1-9	TTTCTACTAAGTGTAGATCCTGCTGGATGTCCTCCCGCGAATCAATTTCTACTAAGTGTA
Cas12a-MGA1-10	TTTCTACTAAGTGTAGATGGCGCCGGAGGCTGTGTGGCGACGGAATTTCTACTAAGTGTA
Cas12a-CAN1-1	TTTCTACTAAGTGTAGATCTACCCGATATCTGTCACAGTCGTTAATTTCTACTAAGTGTA
Cas12a-CAN1-2	TTTCTACTAAGTGTAGATACGACCCCAAGCTGACCGATGACTCAATTTCTACTAAGTGTA
Cas12a-CAN1-3	TTTCTACTAAGTGTAGATGGCAGGAAACTCCAACGTCTACATTAATTTCTACTAAGTGTA
Cas12a-CAN1-4	TTTCTACTAAGTGTAGATGTCTGCTGGCCTTCATGTCTGTGTCAATTTCTACTAAGTGTA
Cas12a-CAN1-5	TTTCTACTAAGTGTAGATGTGCCTCCATGGGCTGGCTATACTGAATTTCTACTAAGTGTA
Cas12a-CAN1-6	TTTCTACTAAGTGTAGATCATCTTCTACATTGGCTCTATCTTCAATTTCTACTAAGTGTA
Cas12a-CAN1-7	TTTCTACTAAGTGTAGATTGGGGTTCTGGGCCTCACCGGCAGTAATTTCTACTAAGTGTA
Cas12a-CAN1-8	TTTCTACTAAGTGTAGATCTTGTGCGAGGGCACCTCCTCTGAGTTTTTTACGTCTAAGAA
Cas12a-CAN1-9	TTTCTACTAAGTGTAGATGTGCGGTTCCGGAGTCAGCCAGGGCAATTTCTACTAAGTGTA
Cas12a-CAN1-10	TTTCTACTAAGTGTAGATCTCGAATTTGCATCTTCTACATTGGAATTTCTACTAAGTGTA
Cas12a-MFE1-1	TTTCTACTAAGTGTAGATAGAGCCCCACCTACCCTAACGGCCCAATTTCTACTAAGTGTA
Cas12a-MFE1-2	TTTCTACTAAGTGTAGATGCCATGTAACCAGCACCGACCTCGTAATTTCTACTAAGTGTA
Cas12a-MFE1-3	TTTCTACTAAGTGTAGATGGGGGTGACACCCTTCTTGGTGTTGAATTTCTACTAAGTGTA
Cas12a-MFE1-4	TTTCTACTAAGTGTAGATGGTGCCTACAAGGTTACCCGAGCTGAATTTCTACTAAGTGTA

Cas12a-MFE1-5	TTTCTACTAAGTGTAGATATGTCCACCTCAACGGTACTTACTCAATTTCTACTAAGTGTA
Cas12a-MFE1-6	TTTCTACTAAGTGTAGATCCGACTTTCTGGTGATTACAACCCTAATTTCTACTAAGTGTA
Cas12a-MFE1-7	TTTCTACTAAGTGTAGATCGGAAACTTCGGCCAGACCAACTACAATTTCTACTAAGTGTA
Cas12a-MFE1-8	TTTCTACTAAGTGTAGATGGTCGTTTCGCTTCGCTGCGCTTGTAATTTCTACTAAGTGTA
Cas12a-MFE1-9	TTTCTACTAAGTGTAGATAAGAAGTCAGCAGGGCCGTTAGGGTAATTTCTACTAAGTGTA
Cas12a-MFE1-10	TTTCTACTAAGTGTAGATTCCTTCTGTGTGGTGTCGTTTTGGGAATTTCTACTAAGTGTA

Table S4.9. Transformation efficiencies measured as x10^6 transformants, for all replicates in the control and treatment strains.

Strain	***Replicate Transformation Efficiency (x10^6 transformants)***		
	R1	**R2**	**R3**
PO1f *Δku70*	689	621	543
PO1f Cas12a *Δku70*	506	429	441

Table S4.10. Primers used for NGS fragment amplification

Primer name	Primer Sequence	Illumina Barcode (Reverse primer) / Pseudo-Barcode (Forward primer) for demultiplexing
ILU1-F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTTCCGGGTCGGCGCAAATTTC	^TTCCGG
ILU2-F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGATCGGGTCGGCGCAAATTTCT	^AGATCG
ILU3-F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTATTCGGGTCGGCGCAAATTTCT	^GCTATT
ILU4-F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGGACTACGGGTCGGCGCAAATTTCT	^CAGGAC
ILU1-R	CAAGCAGAAGACGGCATACGAGATTCGCCTTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	CAAGGCGA
ILU2-R	CAAGCAGAAGACGGCATACGAGATGACGAGAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	CTCTCGTC
ILU3-R	CAAGCAGAAGACGGCATACGAGATAGACTTGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	CCAAGTCT
ILU4-R	CAAGCAGAAGACGGCATACGAGATCTGTATTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	TAATACAG
ILU5-R	CAAGCAGAAGACGGCATACGAGATCCTGAACCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	GGTTCAGG
ILU6-R	CAAGCAGAAGACGGCATACGAGATATCAGGTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	AACCTGAT
ILU7-R	CAAGCAGAAGACGGCATACGAGATTAGGTGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	GTCACCTA
ILU8-R	CAAGCAGAAGACGGCATACGAGATCGAACAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	ACTGTTCG
ILU9-R	CAAGCAGAAGACGGCATACGAGATGTTCGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	GATCGAAC
ILU10-R	CAAGCAGAAGACGGCATACGAGATACCTAGCTGTGACTGGAGTTCAGACGTGTGCCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	AGCTAGGT
ILU11-R	CAAGCAGAAGACGGCATACGAGATAGAGATGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	TCATCTCT
ILU12-R	CAAGCAGAAGACGGCATACGAGATCTGGACTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	AAGTCCAG

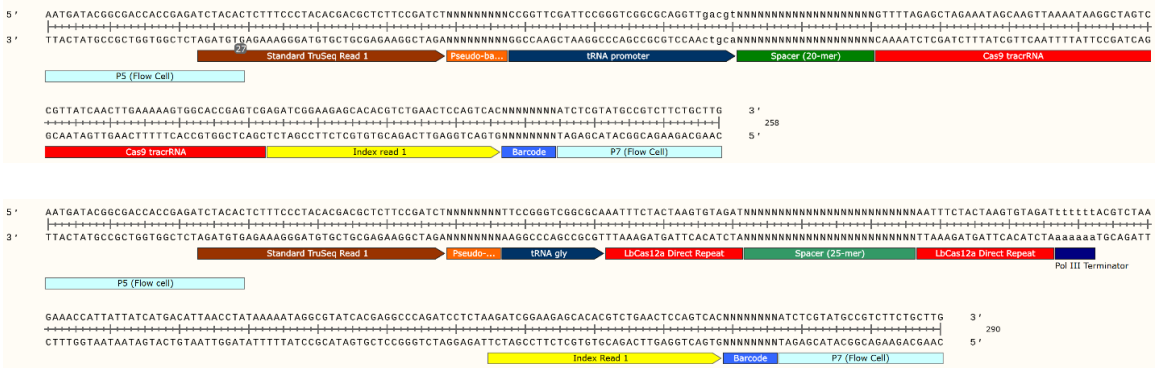

Figure S4.8. Schematic and sequence information of Cas9 (top) and Cas12a (bottom) amplicons for NGS. Amplicons contain (i) P5 and P7 sequences (light blue) that are necessary for binding with the flow cell in Illumina sequencers, (ii) TruSeq adapter (brown) for binding of the sequencing primer, (iii) a portion of $tRNA^{gly}$ (black) expressing the sgRNA, (iv) Cas9 or Cas12 spacer (green) (v) Cas12a associated direct repeats or a portion of the Cas9 tracrRNA sequence (red), (vi) Universal 8 bp Illumina barcodes (blue), (vii) Index read 1 sequence for the binding of primers to sequence the Illumina barcodes, and (viii) 4-9 nt pseudo-barcodes (orange) at the 5' end between the TruSeq and $tRNA^{gly}$ which help demultiplex replicates that contain the same illumine barcode.

Table S4.11. Parameters for bioinformatics tools used in analysis of NGS reads

Tool	Version	Parameters*
FastQC	v0.11.8	Default settings
Cutadapt	Galaxy Version 1.16.6 [8]	The 3 biological replicates of a given sample at a given time-point always had the same reverse primer containing the Illumina barcode, and forward primers ILU1-F, ILU3-F and ILU4-F; or ILU2-F, ILU3-F and ILU4-F each containing different pseudo-barcodes. Thus Cutadapt was used to demultiplex biological replicates from each other. • 5' (Front) anchored 6 bp pseudo-barcodes to be demultiplexed (-g): ^NNNNNN (refer to previous table for pseudo-barcode-forward primer association). • Maximum error rate (--error-rate): 0.2 • Match times (--times): 1 • Minimum overlap length (--overlap): 4 • Multiple output: Yes (Each demultiplexed readset is written to a separate file)
Trimmomatic	v0.38	• HEADCROP: 29 (if amplified by ILU1-F); or 31 (if amplified by ILU2-F); or 32 (if amplified by ILU3-F); or 34 (if amplified by ILU4-F) • CROP: 25
Bowtie2	v2.4.2	• Number of allowed mismatches in seed alignment (-N): 1 • Length of the seed substring (-L): 21 • Function governing interval between seed substrings in multiseed alignment (-i): S,1,0.50 • Function governing maximum number of ambiguous characters (--n-ceil): L,0,0.15 • Alignment mode: end-to-end • Number of attempts of consecutive seed extension events (-D): 20 • Number of times re-seeding occurs for repetitive reads: 3 • Save mapping statistics: Yes

Note: All parameters other than those mentioned here are kept at default values.

Table S4.12. Correlation of SRA files names to demultiplexing information

SRA file name	SRA sample name	Demultiplexing needed	Pseudo-Barcode for Demultiplexing with CutAdapt**	Readsets contained
GW-Cpf1_Control-2_S2_R1_001.fastq.gz	PO1f_dku70_day2_All3reps	Yes	^AGATCG	Replicate #1
			^GCTATT	Replicate #2
			^CAGGAC	Replicate #3
GW-Cpf1_Control-4_S4_R1_001.fastq.gz	PO1f_dku70_day4_All3reps	Yes	^AGATCG	Replicate #1
			^GCTATT	Replicate #2
			^CAGGAC	Replicate #3
GW-Cpf1_Control-6_S6_R1_001.fastq.gz	PO1f_dku70_day6_All3reps	Yes	^AGATCG	Replicate #1
			^GCTATT	Replicate #2
			^CAGGAC	Replicate #3
Yl-Cpf1_CS-2_S2_R1_001.fastq.gz	PO1f_LbCas12a_dku70_day2_All3reps	Yes	^AGATCG	Replicate #1
			^GCTATT	Replicate #2
			^CAGGAC	Replicate #3
	PO1f_LbCas12a_dku70_day4_All3reps	Yes	^AGATCG	Replicate #1

Yl-Cpf1_CS-4_S4_R1_001.fastq.gz			^GCTATT	Replicate #2
			^CAGGAC	Replicate #3
Yl-Cpf1_CS-6_S6_R1_001.fastq.gz	PO1f_LbCas12a_dku70_day6_All 3reps	Yes	^AGATCG	Replicate #1
			^GCTATT	Replicate #2
			^CAGGAC	Replicate #3
GW_Yl_Cpf1-7_S7_R1_001.fastq.gz	LbCas12a_Library_Rep1	No	N/A	Replicate #1
GW_Yl_Cpf1-8_S8_R1_001.fastq.gz	LbCas12a_Library_Rep2	No	N/A	Replicate #2
GW_Yl_Cpf1-9_S9_R1_001.fastq.gz	LbCas12a_Library_Rep3	No	N/A	Replicate #3

** The symbol '^' before the barcode sequence represents that it is anchored, i.e the read begins with the barcode sequence from the 5' end. This information is needed for demultiplexing with CutAdapt.

References

1. Schwartz, C. *et al.* Validating genome-wide CRISPR-Cas9 function improves screening in the oleaginous yeast Yarrowia lipolytica. *Metab. Eng.* **55**, 102–110 (2019).

2. Guo, J. *et al.* Improved sgRNA design in bacteria via genome-wide activity profiling. *Nucleic Acids Res.* **46**, 7052–7069 (2018).

3. Wang, D. *et al.* Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* **10**, 4284 (2019).

4. Moreb, E. A. & Lynch, M. D. Genome dependent Cas9/gRNA search time underlies sequence dependent gRNA activity. *Nat. Commun.* **12**, 5034 (2021).

5. Kim, H. K. *et al.* Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nat. Biotechnol.* **36**, 239–241 (2018).

6. Ramesh, A., Ong, T., Garcia, J. A., Adams, J. & Wheeldon, I. Guide RNA Engineering Enables Dual Purpose CRISPR-Cpf1 for Simultaneous Gene Editing and Gene Regulation in Yarrowia lipolytica. *ACS Synthetic Biology* vol. 9 967–971 (2020).

7. Schwartz, C. M., Hussain, M. S., Blenner, M. & Wheeldon, I. Synthetic RNA Polymerase III Promoters Facilitate High-Efficiency CRISPR-Cas9-Mediated Genome Editing in Yarrowia lipolytica. *ACS Synth. Biol.* **5**, 356–359 (2016).

8. Jalili, V. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.* **48**, W395–W402 (2020).

Chapter 5: Improving the accuracy of functional genomic CRISPR screens in the yeast *Yarrowia lipolytica*

5.1 Abstract

High throughput CRISPR screens are revolutionizing the way scientists unravel the genetic underpinnings of novel and evolved phenotypes. One of the critical challenges in accurately assessing screening outcomes is accounting for the variability in sgRNA cutting efficiency. Poorly active guides targeting genes essential to screening conditions obscure the growth defects that are expected from disrupting them. In this chapter, we address this problem in two ways. (i) We develop acCRISPR, an end-to-end pipeline that identifies essential genes in pooled CRISPR screens using sgRNA read counts obtained from next-generation sequencing. acCRISPR uses experimentally determined cutting efficiencies for each guide in the library to provide an activity correction to the screening outcomes, thus determining the fitness effect of disrupted genes. This is accomplished by calculating an optimization metric that quantifies the tradeoff between guide activity and library coverage, which is maximized to accurately classify genes essential to screening conditions. CRISPR-Cas9 and -Cas12a screens were carried out in the non-conventional oleaginous yeast *Yarrowia lipolytica* to determine a high-confidence (consensus) set of essential genes for growth under glucose, a common carbon source used for the industrial production of oleochemicals. acCRISPR was also used in screens quantifying relative cellular fitness under high salt and low pH conditions to identify known and novel genes that were related to stress tolerance. (ii) Learnings from the reported Cas9 screen and the sgRNA activity prediction algorithm DeepGuide, were leveraged to design a second Cas9

library spanning every ORF at a 3-fold coverage, only consisting of known or predicted high activity sgRNA. Once again, the goal of this design was intended to limit the effect of poorly active sgRNA on determining gene fitness effects, with the added advantage of limiting the transformation efficiency burden inherent in pooled screens, by reducing library size. Essential genes for growth under glucose identified using the optimized library had a high confidence of also belonging to the consensus set. The optimized library developed here will prove useful in conducting accurate functional genetic screens in *Y. lipolytica* strains of diverse backgrounds. Collectively, this work presents an experimental-computational framework for CRISPR-based functional genomics studies that may be expanded to other non-conventional organisms of interest.

This chapter has been submitted as an article to *bioRxiv*. The original citation is as follows: Ramesh, A., Trivedi, V., Schwartz, C., Tafrishi, A., Mohseni, A., Li, M., Lonardi, S., & Wheeldon, I. (2022). acCRISPR: An activity-correction method for improving the accuracy of CRISPR screens. *bioRxiv*.

5.2 Introduction

Functional genetic screening with pooled libraries of CRISPR guides has been successful in discovering gene function, identifying essential genes, and evolving new phenotypes [1–3]. These screens work by inducing mutations across the genome to disrupt gene function. Genome-wide transcriptional regulation is also possible when a catalytically deactivated Cas endonuclease (typically, Cas9 or Cas12a) fused to an activation or repression domain is targeted to promoters [4,5]. For these screens to be effective, the library should contain one or more active guide RNAs for each targeted gene. Creating such libraries is challenging due to imperfect design algorithms and an incomplete understanding of how Cas endonucleases function across different species. Further confounding guide design is the blocking effect of chromatin structure on guide RNA targeted Cas9 endonuclease [6,7]. As a result of this imperfect design, CRISPR screens are conducted with pooled libraries of guide RNAs that have a broad range of activity [8,9]. High activity guides can assign phenotypic changes to genome edits with high confidence, while inactive and low activity guides can obscure gene hits by producing false negatives. Computational and experimental methods that can quantify the activity of each guide in a library and account for the variance in activity are needed to correct screening outcomes, accurately identify genotype-phenotype relationships, and call essential genes with high confidence.

A common CRISPR library design strategy is to include many guides targeting each gene or promoter. This strategy helps ensure that every gene is targeted by an active guide, but doing so increases the analytical complexity in assessing outcomes. Ideally, having a

smaller library that only consists of highly active guides targeting any given ORF will increase confidence in screening outcomes while also relieving the transformation efficiency requirements for robust CRISPR screens [1,42]. However, this would require species specific sgRNA design algorithms that are capable of predicting highly active sgRNA with high accuracy.

Current analysis methods use a Bayesian framework to infer guide activity from screens obtained across several experimental conditions; guide RNAs that elicit a fitness effect under several different conditions are indicative of high activity [10,11]. Reliable measurements of guide activity can also be generated directly from screening experiments. In the yeast species that we have studied[12], this can be achieved by disrupting the primary DNA repair mechanism (typically, non-homologous end-joining or NHEJ) and using negative growth selections to quantify the activity of each guide, resulting in activity profiles across the genome. Guide activity data, whether computationally or experimentally produced, is used to identify and account for inactive and low activity guides, leading to improved hit calling and screen accuracy. Here we show that, given experimental guide activity measurements from a single screen, significant hits can be identified using average log_2-fold change, thereby eliminating the need to process multiple screens and perform probabilistic modeling of the data.

In this work, we tackle the ramifications associated with poorly performing sgRNA within pooled CRISPR screens in two ways. First, we develop an activity-correction CRISPR screen analysis method – acCRISPR – that optimizes library activity to generate

accurate screening outcomes. Using guide RNA abundance data from sample and control screens along with information on the activity of each guide, acCRISPR computes a fitness score for every targeted gene and identifies genes essential to the screening condition. We demonstrate the utility of acCRISPR by analyzing CRISPR-Cas9 and -Cas12a screens in both positive and negative selection experiments in the oleaginous yeast *Yarrowia lipolytica*. We focus on this yeast because it has the ability to synthesize and accumulate lipids, and for its success as a host for oleochemical biosynthesis [13–15]. Using previously derived guide activity profiles of *Yarrowia* genome-wide Cas9 and -12a libraries (see ref. [16]), along with new growth screens, we use acCRISPR to identify essential genes and call hits in low pH and high salt growth screens. We also evaluate the performance of acCRISPR when with computational predictions of guide activity rather than experimentally determined values. Secondly, we design a second Cas9 library consisting only of known or predicted highly active sgRNA targeting all protein coding genes at a 3-fold coverage (half the coverage of the first Cas9 library). We use this library again to determine essential genes and compare them to the set identified by acCRISPR analysis of the first library. Essential gene analysis and functional genetic screening will help toward developing a better understanding of *Yarrowia*'s genetics, and acCRISPR analysis of the screens conducted in this work enables this.

5.3 Results

5.3.1 acCRISPR optimizes sgRNA library activity and coverage.

acCRISPR uses raw read counts of guide RNAs from functional screens as inputs and computes cell fitness effects, guide RNA activity profiles, and calls essential genes. To

demonstrate this analysis pipeline, we conducted CRISPR-Cas9 and -Cas12a genome-wide screens in the PO1f strain of *Y. lipolytica.* The pooled guide libraries contain single guide RNAs (sgRNAs) that target more than 98.5% of the protein-coding sequences with 6- and 8-fold coverage for Cas9 and Cas12a, respectively. Guide activity in these libraries was previously reported [9,16]; a cutting score (CS), defined as the $-log_2$ ratio of normalized read counts obtained in PO1f Cas9/12a *ΔKU70* to counts in the control strain, was determined for each guide (**Fig. 1a**). The disruption of *KU70* disables NHEJ DNA repair [17], creating a link between guide abundance in a negative selection growth screen and guide activity. In the absence of the dominant DNA repair mechanism, a double-stranded break causes cell death or significant impairment in growth; sgRNAs with high activity are lost from the cell population with higher frequency than those with lower activity, thus linking CS to guide activity. The fitness screen inputs for acCRISPR were generated using PO1f as the control strain and PO1f Cas9 or Cas12a as the sample. Screens were conducted in synthetic defined media with glucose as the sole carbon source. An Illumina sequencing instrument was used to generate sgRNA read counts after four days of culture. These data were used to generate a fitness score (FS) profile, defined as the log_2 ratio between the normalized counts in the Cas9/Cas12a expressing strain and the control. Raw guide RNA counts for Cas9 and Cas12a screens are provided in **Supplementary Files 5.1 and 5.2**.

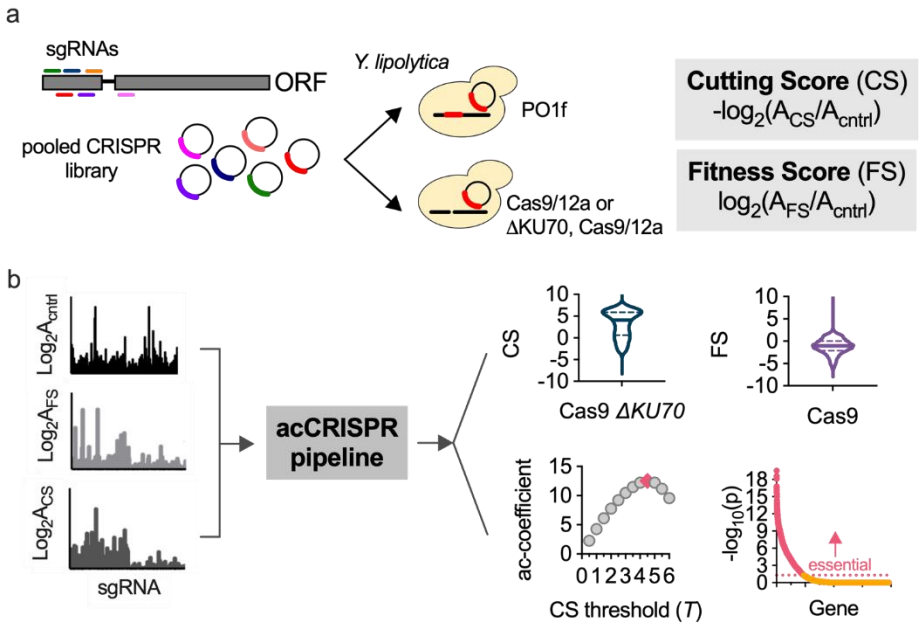

Figure 5.1. acCRISPR analysis of CRISPR-Cas screens. (**a**) Growth screens in *Y. lipolytica* were conducted with pooled libraries of single guide RNAs (sgRNAs) (6- and 8-fold coverage of >98.5% of CDSs, for Cas9 and Cas12a respectively). A guide's cutting score (CS) is equal to the *-log*$_2$ fold-change of normalized guide abundance in PO1f Cas9/12a *ΔKU70* to the control strain. Fitness scores (FS) are similarly defined, but with the PO1f Cas9/12a strain as the sample. (**b**) acCRISPR takes normalized sgRNA read counts from the control, CS, and FS strains and computes a series of outputs: CS per guide, FS per gene, the ac-coefficient (the product of $CS_{threshold}$ and library coverage), and p-value per gene from significance testing against a non-essential gene population at the maximum ac-coefficient. The data sets shown here are from Cas9 screens in *Y. lipolytica* PO1f. Screens were conducted at 30 °C with glucose as the sole carbon source. Genes with an essentiality p-value <0.05 were classified as essential.

The first analytical step of acCRISPR is to convert raw guide abundance values into CS and FS profiles (**Fig. 5.1b, Supplementary File 5.3**). First, an FS is computed for each gene as the average *log*$_2$-fold change of all guides targeting that gene, both active and inactive. Then, the FS value for each gene is recalculated after excluding sgRNAs with a CS below a given CS threshold (*i.e.,* a minimum value of CS for an sgRNA to be included

in the analysis, *T*). As guides with low CS are removed, the library coverage is reduced along with the statistical power that multiple guides provide. To capture this effect, we compute the ac-coefficient as the product of the CS threshold (*T*) and the average number of guides per gene, for a range of *T* values. The maximum peak for the ac-coefficient indicates the CS threshold where the library activity is maximized. The corrected FS profile generated for the threshold corresponding to the peak is used to identify essential gene hits; p-values for every gene in the dataset are determined by comparing the FS of a gene to a null distribution that represents the fitness of non-essential genes (see Methods for more details).

5.3.2 acCRISPR accurately calls essential genes.

We evaluated the performance of acCRISPR against other established approaches that classify essential genes using read counts or log_2-fold changes from CRISPR screens as input, namely JACKS [10], MAGeCK-MLE [11], and CRISPhieRmix [18]. These methods have been validated against a gold standard set of essential genes in mammalian cells and were used here to compute fitness effects and call essential genes in *Yarrowia*. The comparison of acCRISPR to the other methods on our Cas9 screens is shown in **Fig. 5.2**. Similar analyses of the CRISPR-Cas12a screens are shown in **Supplementary Fig. S5.1**.

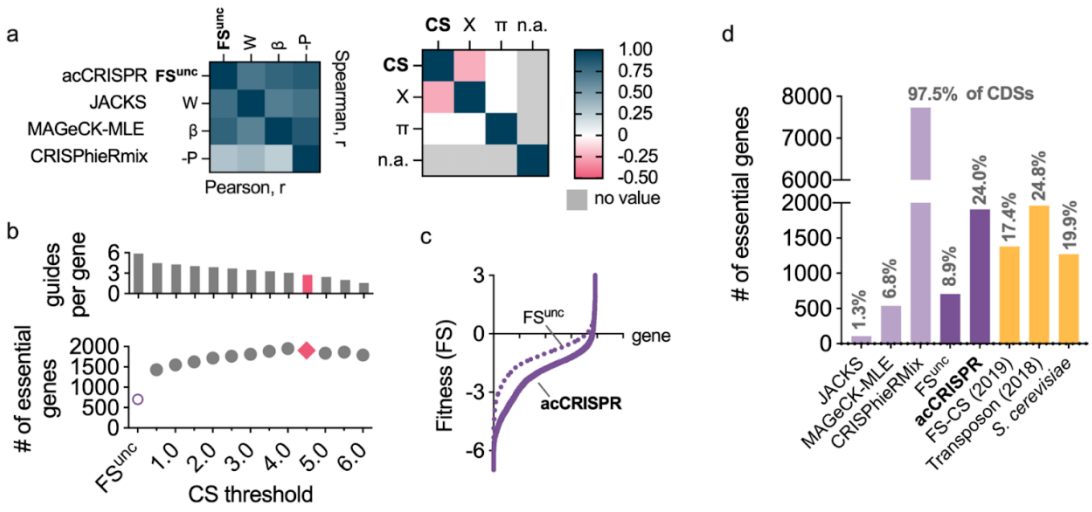

Figure 5.2. acCRISPR analysis of CRISPR-Cas9 screens defines a high confidence set of essential genes. (**a**) Heat maps showing Pearson (below diagonal) and Spearman (above diagonal) correlation coefficients for comparison of gene fitness effects (uncorrected FS (FS^{unc}), W, β, and -P; left) and sgRNA cutting efficiencies (CS, X, and π; right) from acCRISPR and three established essential gene identification algorithms, JACKS, MAGeCK-MLE and CRISPhieRmix. 'n.a.' denotes that sgRNA cutting efficiency values for CRISPhieRmix are not available. (**b**) The average number of sgRNAs per gene and the number of essential genes predicted with increasing CS threshold (bottom). The number of essential genes predicted for the corrected and uncorrected analyses. The data points colored in pink are the guides per gene and the number of essential genes determined at the maximum ac-coefficient. (**c**) Fitness scores of genes with (solid line) and without (dashed line) acCRISPR processing with a CS threshold (*T*) of 4.5. (**d**) The number of essential genes identified by JACKS, MAGeCK-MLE, CRISPhieRmix, FS^{unc}, and acCRISPR are compared to previously reported essential gene sets for *Yarrowia* (FS-CS[8] and transposon analysis [19]) and *S. cerevisiae* [20]. Values at the top of each bar indicate the percentage of the total number of genes identified as essential by the respective method.

acCRISPR, JACKS, and MAGeCK-MLE output values for the fitness effect of genes in *Yarrowia* (FS uncorrected (FS^{unc}), W, and β) are in good agreement. The pairwise Pearson and Spearman r-values are 0.65 or greater (**Fig. 2a**). CRISPhieRmix was less successful at

capturing raw fitness effects from the *Yarrowia* screen (Pearson r <0.37) and the majority of genes were identified as essential. JACKS and MAGeCK-MLE also output guide activity predictions (X and π); these values did not correlate well with the acCRISPR analysis of the CS profiles, which were directly obtained from the screening experiment.

We next applied CS correction to the Cas9 screening data. The ac-coefficient curve for the Cas9 screen for each choice of the CS threshold *T* is shown in **Fig. 5.1b**. The number of essential genes and the average number of guides per gene for the same values of the threshold *T* are shown in **Fig. 5.2b**. As *T* increased from 0.5 to 4.0, the number of genes classified as essential also increased, an effect likely caused by removing false negatives resulting from poor activity sgRNAs targeting essential genes. The optimum library activity, indicated by the peak of the ac-coefficient, occurred at threshold *T*=4.5 with an average coverage of 2.78 guides per gene. The peak for the ac-coefficient in the CRISPR-Cas12a library indicated the optimal CS threshold of *T*=1.5, with an average coverage of 2.97 guides per gene (**Supplementary Fig. S5.1**).

The optimized acCRISPR analysis of the Cas9 screen identified 1903 essential genes (see **Supplementary File 4**), a number similar to the 1954 essential genes reported for a transposon-based screen [19]. Without the activity correction, only 702 genes could be classified as essential, a value significantly below what was expected; based on the analysis of other yeast species ~15% to ~30% of protein-coding genes are expected to be essential (*e.g.*, 19.9% for *S. cerevisiae* and 26.1% for *S. pombe* [20,21]). The Cas12a screens conducted here identified 1375 genes as essential (**Supplementary File 5.4**) when the acCRISPR

pipeline was used, and only 335 when all sgRNAs (both active and inactive) were included in the analysis. JACKS and MAGeCK-MLE also under-predicted the number of essential genes in the Cas9 and Cas12a screens (JACKS, 102 and 0 ; MAGeCK-MLE, 535 and 1218), while CRISPhieRmix classified nearly all genes as essential (7724 and 7538).

5.3.3 CRISPR-Cas9 and -Cas12a screens help define a consensus set of essential genes.

The acCRISPR analysis of the Cas9 and -12a screens provides the opportunity to define a consensus set of essential genes for *Yarrowia* growth on glucose. First, we validated the essential gene set via a Gene Ontology (GO) enrichment analysis [22,23], with the expectation that functional terms known to be essential would be enriched (FDR-corrected $p < 0.05$; see **Supplementary Files 5.5 and 5.6** for all GO and GO-Slim terms pertaining to molecular function (MF), biological process (BP) and cellular component (CC)). As expected, genes involved in transcription, translation, cell cycle regulation, cofactor metabolism, and tRNA metabolic processes showed significantly lower FS values (t-test, $p < 0.05$) compared to the average FS of all genes in both the Cas9 and Cas12a screens. The FS values of genes in these functional groups along with other enriched GOSlim terms are shown in **Fig. 3a**.

A previously published transposon-based screen identified 1954 essential genes [19]. Experimental conditions (2% glucose in SD-Leu media) were consistent with the Cas9 and Cas12a experiments conducted here, thus providing a large data set from which we can

identify a consensus set of essential genes. One thousand six hundred and twelve genes were common to at least two of the three different screens (**Fig. 5.3b** and **Supplementary File 5.7**). Enriched GO-Slim terms in this set were consistent with those expected for essential genes and we consider these genes as the consensus set for *Yarrowia* growth on glucose (see **Supplementary File 8**). The essential genes identified in the consensus set were also compared to known essential genes in *S. cerevisiae* and *S. pombe*. Of these, 824 genes were identified to have homologs in *S. cerevisiae*, of which 54.6% were found to be essential in both species. Seven hundred and eighty-two genes had homologs in *S. pombe* and 60.9% of those were found to be commonly essential between both species (**Supplementary Fig. S5.2**).

5.3.4 acCRISPR can use sgRNA activity predictions as an alternative to CS.

We recognize that generating experimental CS profiles is not always feasible (for example, in organisms for which it is not possible to have NHEJ-deficient screens or in cases where a double stranded break is likely to be repaired by homology directly using a second allele as a template). Thus, we sought to test the performance of acCRISPR using computationally predicted sgRNA activity scores in Yarrowia. Among the large set of guide prediction tools available for Cas9, we selected DeepGuide [16], uCRISPR [24], Designer v1 [25], Designer v2 [26], SSC [27], CRISPRscan [28], and CRISPRspec [29] (**Fig. 5.4** and **Supplementary File 5.9**). For Cas12a, only a few prediction algorithms have been developed, for example, DeepGuide [16] and DeepCpf1 [30], which have been shown to predict sgRNA activities in *Yarrowia* with reasonable accuracy (**Supplementary Fig. S5.3 and Supplementary File 5.10**). Using the predicted activity scores, we implemented

acCRISPR to compute the maximum ac-coefficient (**Supplementary Table S5.1**) and determined a set of predicted essential genes.

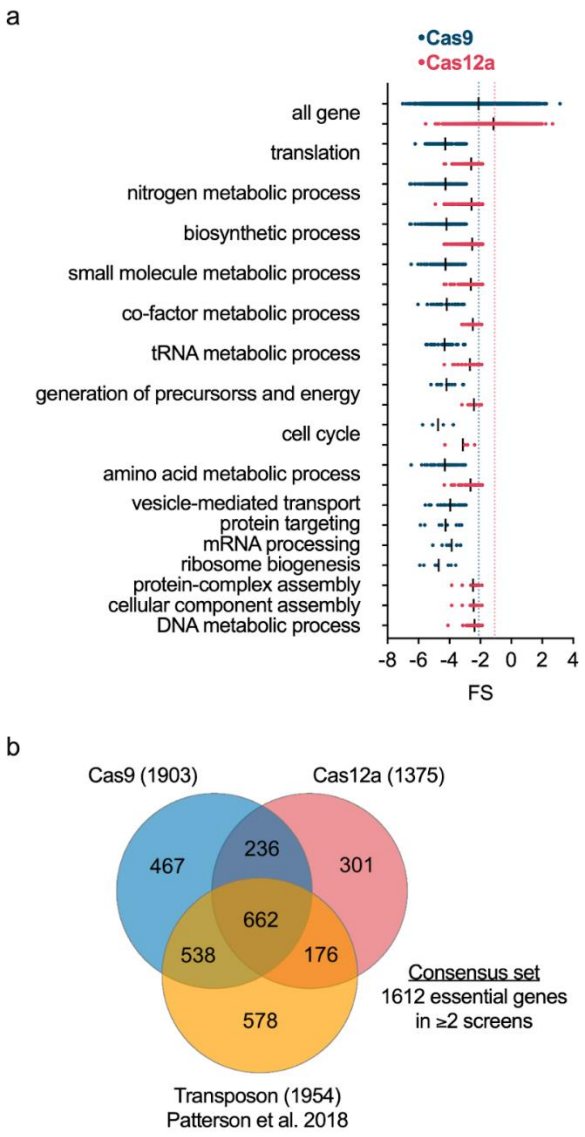

Figure 5.3. Defining a set of consensus essential genes in *Y. lipolytica*. (**a**) Enriched GO-Slim biological process terms for Cas9 and Cas12a essential gene sets and FS distribution of essential genes associated with each GO-Slim term. Enriched terms were determined using a hypergeometric test (FDR-corrected, $p < 0.05$). The FS values for each GO-Slim term were found to be significantly lower than those of all genes by unpaired t-test ($p < 0.0001$). Blue and red dotted lines indicate the mean FS of all genes for Cas9 and Cas12a datasets respectively. (**b**) Venn diagram of the essential genes identified from CRISPR-Cas9, CRISPR-Cas12a, and transposon screening, and their overlap. The consensus set of essential genes, comprising genes common to at least two of the three screens, contains 1612 unique genes.

The consensus set identified in **Fig. 5.3** served as a reference to evaluate the success of each prediction method. Of all prediction methods, DeepGuide was found to have the

highest sensitivity for both Cas9 (62.8%) and Cas12a (51.7%) datasets (where sensitivity is the percentage of the consensus set that is captured by the predicted set). The higher performance of DeepGuide is likely a consequence of its training set, that is the *Yarrowia* CS profiles generated in our screens. Other methods captured a smaller fraction of the consensus set, with sensitivity ranging from 26.0% to 44.9%. While the predicted guide activities were not successful at capturing the full set of essential genes in *Yarrowia*, those that were identified were called with high confidence; each of the tested methods maintained precision rates above ~75% (where precision is the number of predicted essential genes overlapping with the consensus set divided by the total number of essential genes predicted).

In addition to evaluating the success of different guide prediction algorithms, we determined sensitivity and precision metrics for Cas9 and Cas12a screens using acCRISPR, JACKS, MAGeCK-MLE, CRISPhieRmix, and uncorrected FS profiles, with CS as an input (**Fig. 5.4** and **Supplementary Fig. S5.3**). acCRISPR analysis of the Cas9 screen captured nearly all of the consensus set (sensitivity of 89.1%) with high precision (75.5%). Except for CRISPhieRmix, the other methods failed to capture the majority of the consensus set. CRISPhieRmix classified nearly all *Yarrowia* genes as essential, thus capturing nearly 100% of the consensus set but with low precision (20.8%). Results of a similar analysis with the Cas12a screen are reported in **Supplementary Fig S5.3**; the Cas12a screen captured 66.7% of the consensus set with 78.1% precision.

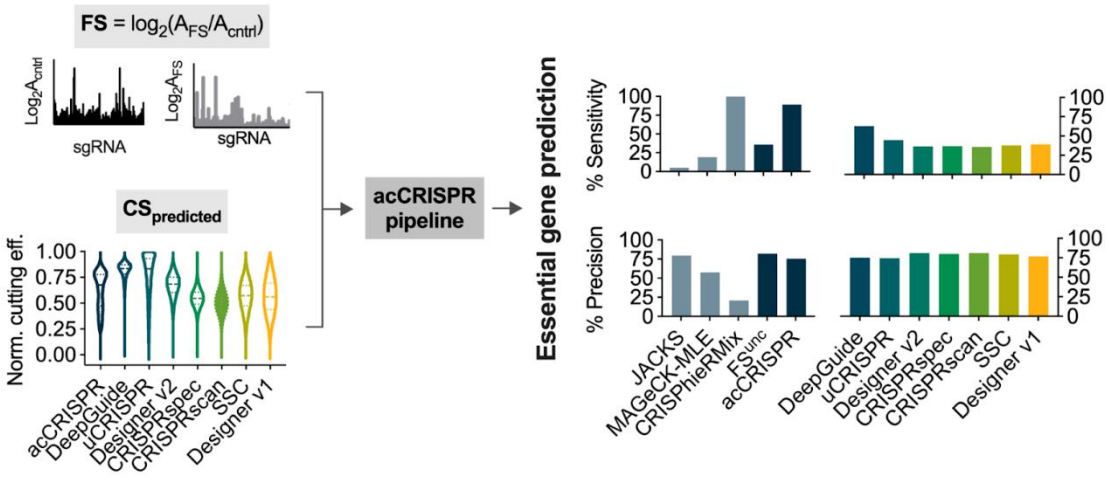

Figure 5.4. Performance of acCRISPR using predicted sgRNA activity profiles in *Y. lipolytica*. Raw sgRNA counts from control and treatment strains used for fitness score calculations were provided as input to acCRISPR along with sgRNA activity scores from a range of guide prediction tools (DeepGuide [16], uCRISPR [24], Designer v2 [26], CRISPRspec [29], CRISPRscan [28], Spacer Scoring for CRISPR (SSC) [27] and Designer v1 [25] left). The violin plot shows the distribution of min-max normalized CS (denoted by 'acCRISPR') and sgRNA activity scores from each prediction tool. Dashed lines represent the median of the normalized score and the dotted lines represent the first and third quartiles. Essential genes were identified using predicted sgRNA efficiency scores from each tool after first determining the maximum ac-coefficient. The % sensitivity and % precision in identifying genes from the consensus set are shown (right). Bars indicate the values of these two metrics for each prediction tool as well as for JACKS, MAGeCK-MLE, CRISPhieRmix, uncorrected FS (FS^{unc}), and acCRISPR.

5.3.5 acCRISPR identifies biologically insightful hits related to stress tolerance.

To further demonstrate the utility of acCRISPR, we conducted a series of high salt and low pH tolerance screens from which we identified genetic hits that produced significant effects on cell fitness. Tolerance to high salinity and acidity are industrially beneficial traits that can reduce costs associated with process sterilization [31]. Salt tolerance can also enable growth in lower-cost water sources (*e.g.*, seawater or wastewater), and the ability to grow

in low pH (*e.g.*, pH 2-3) can benefit lipid accumulation in oleaginous yeasts [32]. The CRISPR-Cas9 strain was grown in the presence and absence of various stress conditions (pH 2.5 and 3, and [NaCl] of 0.75 and 1.5 M) and acCRISPR was used to identify significant hits for each stress condition. As a control, the Cas9-containing strain was grown under standard growth conditions (initial pH 5.8 and no added NaCl). In place of FS, these screens defined a tolerance score (TS), which is equal to the log_2 ratio of sgRNA abundance under the stress condition to that grown under control conditions (**Fig. 5a**). A high TS indicated that gene disruption conferred a growth advantage under the applied stress and vice-versa (see **Supplementary Fig. S5.4** for corrected TS profiles in tolerance screens conducted at 1.5 M NaCl and pH 2.5).

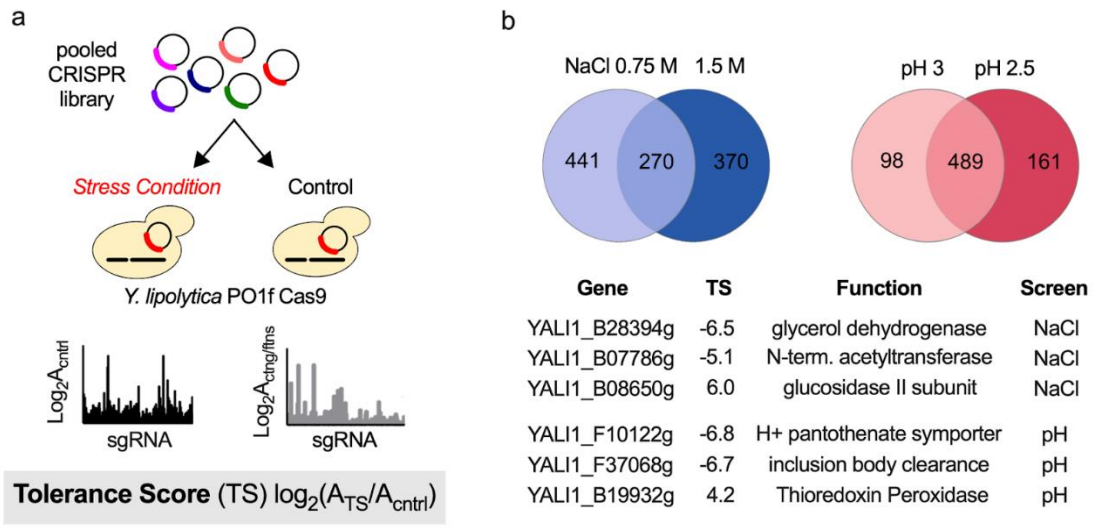

Figure 5.5. acCRISPR analysis of environmental stress tolerance screens. (**a**) Schematic of the CRISPR-Cas9 stress tolerance screens in *Yarrowia*. Analogous to fitness score (FS), the tolerance score (TS) is used to define the effect of each guide on cell growth under a stress condition. TS is equal to the log_2-fold change of sgRNA abundance in the treatment to the control, where the control is a Cas9-expressing strain grown under standard culture conditions. (**b**) Outcomes of high salt and low pH screens. Venn diagrams (top) show the overlap of gene hits identified in the salt (0.75 M and 1.5 M NaCl) and low pH (pH 3 and 2.5) screens. Selected hits are shown (bottom), including the gene ID, the TS value from the 1.5 M NaCl and pH 2.5 conditions, and putative gene function.

acCRISPR analysis of the salt tolerance screens (**Supplementary Fig. S5.5**) identified 270 gene hits that were common to both stress levels (0.75 M and 1.5 M NaCl); 210 of these showed reduced fitness, while the other 60 resulted in increased salt tolerance (**Fig. 5.5b and Supplementary File 5.11**). The top two hits with fitness defects and the top hit with a fitness benefit provided confidence in the screening outcomes as these genes are known to affect salt tolerance in other species. Glycerol dehydrogenase (*GCY1*; TS of -6.5 at 1.5 M NaCl) is directly related to glycerol biosynthesis, which is known to play an important role in hyperosmotic stress resistance [33,34]. Gcy1 protein abundance has also been

shown to increase during DNA replication stress in *S. cerevisiae*, a downstream effect of environmental stress [35]. The second loss-of-fitness hit, *ARD1* (N-terminal acetyltransferase; TS of -5.1 at 1.5 M NaCl), has also been shown to have increased expression under DNA replication stress [36]. Lastly, the top hit that conferred a fitness advantage, *ROT2* (TS of 5.9 at 1.5 M NaCl) is responsible for regulating the chitin composition of the cell wall, and its disruption in *S. cerevisiae* increases chitin, an effect that has been linked to salt tolerance in yeast and plants [37–39].

The low pH screens also yielded several hits that are known to affect acid tolerance (489 hits common to both screens, including 256 that decreased pH tolerance and 233 that increased it). Functional disruption of the *S. cerevisiae* homolog of the top loss-of-fitness hit (TS of -6.8 at pH 2.5), *FEN2* an H+ pantothenate symporter, has been shown to reduce resistance to low pH [40]. The second top hit *IML2* (TS of -6.7 at pH 2.5) produces a protein required for inclusion body clearance and protein abundance is upregulated under DNA replication and protein misfolding stress, a response that is expected in low pH cultures. Lastly, thioredoxin peroxidase (*TSA1*), a top gain-of-fitness hit (TS of 4.2 at pH 2.5), is known to be involved in acidic pH tolerance in *S. cerevisiae*; the null mutant increases growth tolerance to low pH sodium citrate media [41]. The results reported here support the validity of our acCRISPR analysis in identifying novel gene hits related to stress tolerance; the full list of hits will enable us to identify new cellular functions related to stress tolerance as well as identify mutational targets for engineering new strains with increased tolerance.

5.3.6 The optimized minimal Cas9 library consists of highly active sgRNA

Pooled screens require transformation efficiencies of at least 100 times the library size to ensure high statistical confidence in screening results. Thus, a smaller library containing only a core set of highly active guides for each gene target promises to reduce transformation burden, while increasing screening accuracy [1,42]. To that end, we designed a Cas9 library containing only highly active sgRNA, at a 3-fold gene coverage in *Y. lipolytica*. This library was designed leveraging both sgRNA CS values from the prior Cas9 screen, as well as the *Y. lipolytica* specific guide activity prediction algorithm, DeepGuide. For experimentally validated sgRNA, the old Cas9 library was first filtered for sgRNA with CS values greater than 4.0 (max. ac-coefficient threshold for essential gene calling) and picked top two best performing sgRNA for each gene whenever possible. The third sgRNA came from DeepGuide's best prediction for that gene. If two experimentally validated were not present for any gene, they were replaced with one of DeepGuide's top predictions, for a total of 3 guides per gene (**Supplementary Figure S5.6**). MNase-Seq was performed on the PO1f strain of *Y. lipolytica* to determine strain specific nucleosome occupancy (see methods), and this information was supplemented to DeepGuide for guide activity predictions. All sgRNA in the final library were verified to unique to minimize off-target nuclease activity. Library statistics such as coverage per gene, fraction of experimental and predicted guides in the final library, as well as the frequency distribution of guides in the final cloned library is shown in **Supplementary Figure S5.6**. All 7919 protein coding genes had 3 sgRNA, and the nearly identical mean and median, and high kurtosis, indicates an even T-distribution of guides in the library.

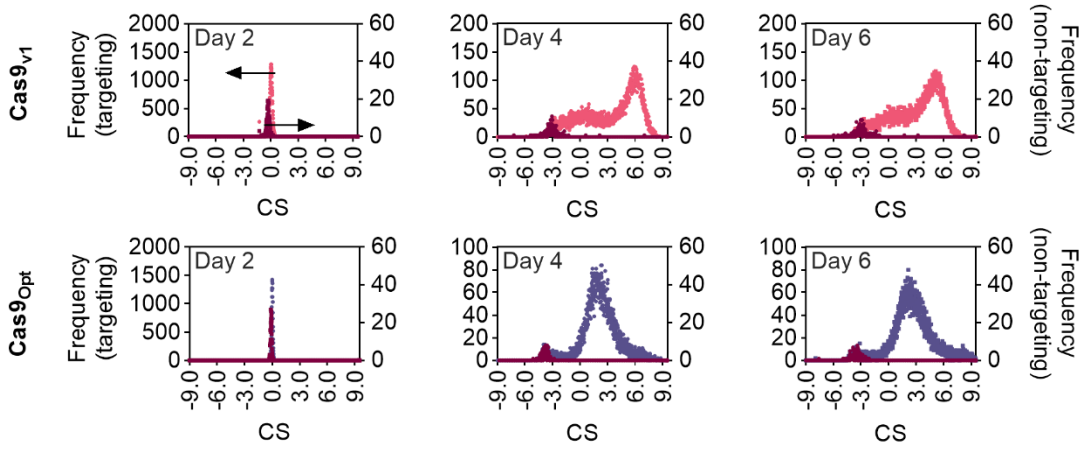

Figure 5.6. Cutting score (CS) distributions of old and optimized Cas9 libraries in *Yarrowia lipolytica*. CS distributions were calculated across three separate days after subculturing transformants twice when they reached confluency. Purple and Pink distributions plotted on the left y-axis show CS values of optimized and the older Cas9 libraries respectively, while the dark red data plotted with the right y-axis depicts the non-cutting control population, constituting ~1.5% of the respective library. The higher the value of CS, the better the cutting activity of the sgRNA. (a) Histogram of CS values in the old unoptimized library. (b) Histogram of CS values in the smaller optimized Cas9 library. The CS values at Day 4 for both libraries were carried forward for further analysis.

The CS distributions of the optimized library determined at two, four and six days of growth are shown in **Figure 5.6**. After only two days of culture, CS values remained close to zero indicating minimal guide activity (at day 2, CS_{avg}=-0.0004). Observed CS profiles at day 4 and day 6 remained unchanged with very similar mean CS values (at day 4, CS_{avg}=2.05; at day 6 CS_{avg}=2.57), thus we elected to use day 4 data as we have shown previously with the Cas9 and Cas12a screens. The library also included non-targeting sgRNA represented at 1.5% that functioned as negative controls. As expected, these guides showed very poor cutting activity at day 4 ($CS_{avg,NT}$=-3.91). When comparing the

distributions of the optimized Cas9 library to the first iteration of the Cas9 library, it was immediately evident that the optimized library had two distinct peaks indicative of highly active and poorly active sgRNA, that were well separated. The targeting guides sgRNA collapsed into a single distribution at a mean of CS=2.14, while in the prior Cas9 library, the targeting population showed a peak and a shoulder, with a mean CS=3.37. Meanwhile, the nontargeting population in the optimized library shows a lower CS than in the first Cas9 library (Optimized library: $CS_{avg,NT}$=-3.91; Cas9 library v1: $CS_{avg,NT}$=-3.07). **Supplementary Figure S5.7** shows the CS distributions of both Cas9 libraries with the mean of the nontargeting populations normalized to 0. A small CS threshold of 1.0, nearly 5 $\log_2$ units from the mean of the non-targeting population (32-fold enrichment in abundance compared to the non-targeting sgRNA) was applied as a qualitative filter to retain highly active sgRNA from this library (**Supplementary Figure S5.8**). We observed that 68.3% of all predicted sgRNA and 85.6% of all experimental sgRNA cleared this threshold, accounting for a total of 80% of total library that showed high activity.

5.3.7 Performance validation of the optimized Cas9 library on essential gene prediction

acCRISPR was used to call essential genes with the fitness scores (FS) obtained on day 4 from the screens conducted with optimized libraries. Essential genes were first predicted without, and with a small activity correction at a CS threshold of 1.0. Once again, the consensus set identified in **Fig. 5.3** served as a reference to evaluate the success of both predicted essential gene sets (**Fig. 5.7b**). With no activity correction, the optimized library only captures only half of the consensus set, with a sensitivity of 55.4%. At a CS of

threshold of 1.0 however, the optimized library captured over 75% of the consensus set, discarding only 20% of library guides as poor cutters. The older Cas9 library showed very poor capacity to capture the consensus set without activity correction (Sensitivity=35.7%), but did capture 89% of the consensus set at an activity correction threshold of 4.0 discarding over 52% of library guides. The precision, which estimates the fraction of predicted genes that belong to the consensus set, was 74.5% and 72.1% respectively for the optimized library without and with activity correction. In comparison, the precision of the older Cas9 library was 82.1% and 75.4% without and with activity correction. While the older unoptimized Cas9 library does capture the consensus set more effectively, it does so at the cost of having more coverage of guides. The optimized library comes close to matching this performance, with less than half the number of total sgRNA, and also discarding fewer sgRNA as nonperforming.

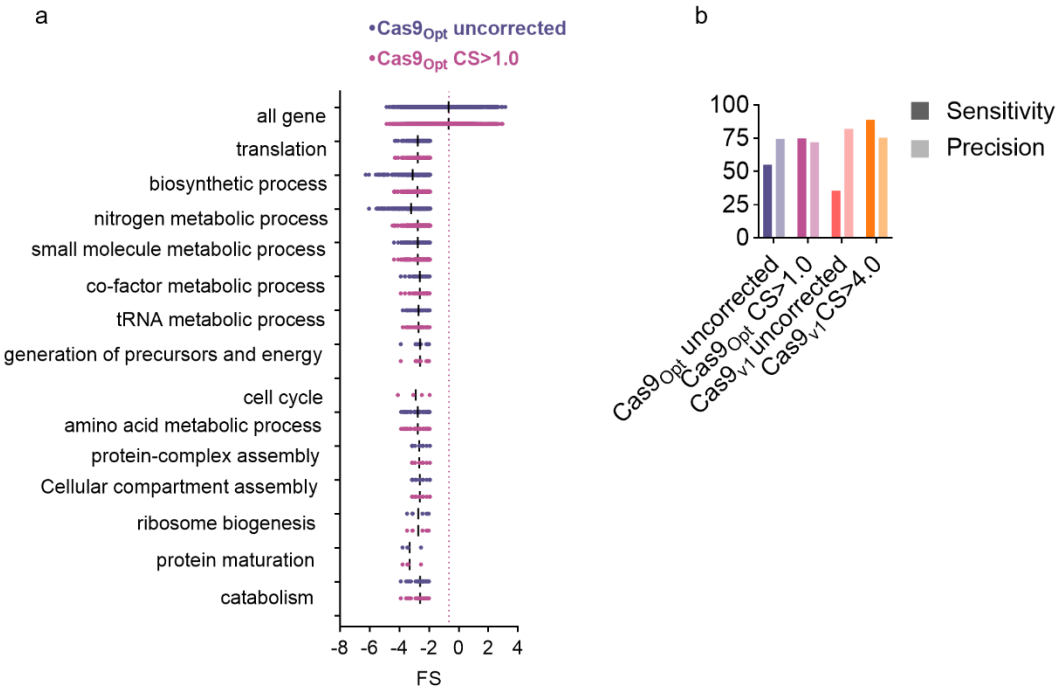

Figure 5.7. Characterization of essential gene sets determined by the optimized Cas9 library. (**a**) Enriched GO-Slim biological process terms for uncorrected and $CS_{thresold}$>1.0 Cas9 essential gene sets and FS distribution of essential genes associated with each GO-Slim term. Enriched terms were determined using a hypergeometric test (FDR-corrected, $p < 0.05$). The FS values for each GO-Slim term were found to be significantly lower than those of all genes by unpaired t-test ($p < 0.0001$). Purple dotted line indicates the mean FS of all genes for the Cas9 library with and without activity correction. (b) Sensitivity and precision of the optimized and unoptimized older Cas9 library, with and without their respective activity correction. Darker bars indicate sensitivity values while the lighter bars indicate precision.

Finally, we also validated the essential gene sets obtained with the optimized library (uncorrected and $CS_{threshold}$>1.0) via a Gene Ontology (GO) enrichment analysis [22,23], with the expectation that functional terms known to be essential would be enriched (FDR-corrected $p < 0.05$). As expected, genes involved in transcription, translation, cell cycle

regulation, cofactor metabolism, ribosome biogenesis, and tRNA metabolic processes showed significantly lower FS values (t-test, $p < 0.05$) compared to the average FS of all genes with the optimized Cas9 library screen **(Fig. 5.7a)**.

5.4 Discussion

A central challenge in analyzing CRISPR screens is deconvoluting the effect of poorly active guides from guides that create genome edits and elicit fitness effects. One approach to solving this challenge is to interrogate each edit in an arrayed format. The physical separation of different genetic perturbations throughout the screen also makes this approach more easily combined with -omics based profiling for further characterization of mutants. However, this requires extensive laboratory automation to achieve the throughputs that are accessible to pooled screens, where one can test the effect of all library mutants in a single culture. On the other hand, pooled screens lack distinct separation between mutants and thus rely on next generation sequencing methods to quantify the effect of genetic perturbations on cell fitness. Thus, deconvoluting the effect of non-performing guides becomes ever more important in this context. acCRISPR addresses this issue in pooled screens by optimizing the screen's ac-coefficient, a parameter that balances the trade-off between guide activity and coverage to maximize the performance of the library. In contrast to existing methods that infer sgRNA activity by modeling multiple screening conditions, acCRISPR uses an experimentally derived measure of guide activity obtained from an additional treatment sample in which DNA repair by NHEJ is disrupted. This additional data enabled acCRISPR to outperform other approaches in determining an accurate set of essential genes.

acCRISPR was developed and validated using CRISPR-Cas9 and -Cas12a screening data to define essential genes in the oleaginous yeast *Y. lipolytica*. The other methods tested here, JACKS, MAGeCK-MLE, and CRISPhieRmix, are most commonly used to analyze the outcomes of mammalian cell CRISPR screens, and were found to be incompatible with our *Yarrowia* data; only a small percentage or all genes were identified as essential. This incompatibility is likely because the overlap between the fitness effect profiles of the non-targeting controls and the active sgRNA population is greater in mammalian cells compared to *Yarrowia* (**Supplementary Fig. S5.9** and **see refs.** [18,42]). CRISPhieRmix, which uses the non-targeting population to form the null distribution, greatly overestimates the number of essential genes in *Yarrowia*, classifying nearly all genes as essential. The relative fitness effects that targeting and non-targeting sgRNAs have may also be harder to resolve in mammalian cells due to alternative splicing, polyploidy, and redundant gene function. acCRISPR, on the other hand, uses sgRNA targeting non-essential genes to construct the null model, thereby making it more adaptive to the *Yarrowia* dataset, and potentially more adaptable to other datasets.

While acCRISPR's use of an experimentally derived CS dataset is empowering, it also increases the technical difficulty of the experiments and is not necessarily accessible in all organisms (*e.g.*, activity profiles across mammalian cell genomes and the genomes of other species have not yet been defined). We also recognize that alternate repair mechanisms could mask CRISPR Cas9/12a cutting. For example, we have previously observed error-prone microhomology mediated end-joining (MMEJ) DNA repair in *Yarrowia* [17]. sgRNA that produce such cases will likely result in negative CS and FS values, indicating that

despite poor guide activity, gene editing still occurred at a rate sufficient to affect cell fitness. Analysis of the CS and FS values per guide reveal that only 1.2% and 2.1% of guides from the Cas9 and Cas12a libraries respectively fit this pattern (see **Supplementary File 5.3**). The primary feature of acCRISPR is to remove guides with low CS, as such the majority of cases where an alternative repair mechanism was active will likely be removed from the final analysis.

The ability to use predicted sgRNA activities in place of experimental activity scores may help address the limitation of requiring an experimental dataset. acCRISPR analysis with predicted activity resulted in high precision but modest sensitivity, thereby capturing a small portion of the essential genes but with high confidence (**Fig. 5.4**). While prediction methods have proven effective at designing active CRISPR sgRNAs, predictive power is still limited to the organism from which the training data was generated [8,16,43]. As better guide design algorithms are developed, we anticipate an improvement in acCRISPR performance in resolving essential genes when using predicted guide activities in place of experimentally derived CS distributions.

acCRISPR analysis of the screens conducted here represents a meaningful step toward understanding *Yarrowia* genetics. Thus far, there have only been a few attempts at classifying essential genes [9,19]. We use the CRISPR-Cas9 and -Cas12a screens conducted here along with the outcomes of a transposon screen conducted under similar conditions (see ref. [19]) to define a consensus set of essential genes for growth on glucose. This set contains 1612 genes that were classified as essential in at least two of the three independent

screens, we consider this the consensus set (**Fig. 5.3b**). While a considerable number of essential genes were called by 2 or 3 of the different technologies, a number of genes were unique to each, likely due to mechanistic differences between the mutagenesis strategies. For example, transposon-based screens have sequence biases for insertions and are known to miss shorter genes [44,45]; the more restrictive PAM of Cas12a leads to lower genome-wide coverage; Cas9 has been shown to have higher rates of off-target effects, which could lead to false predictions; and specific to our experiments, the Cas12a library contains more inactive and low activity guides, thus reducing the number of genes targeted by highly active sgRNAs. Defining a consensus set mitigates these differences as well as other potential issues with functional genomic screens (*e.g.*, plasmid instability) and leads to calling a high confidence set of essential genes – that is, those that were called in more than one screen. GO term enrichment analysis suggests that genes in the consensus set have functions expected to be essential (*e.g.*, genes related to transcription, translation, and cell cycle among others; **Supplementary File 5.8**), while those unique to each method have no enriched functions (**Supplementary File 5.12**).

With respect to the high salt concentration and low pH tolerance screens, acCRISPR analysis also helps to advance our understanding of *Yarrowia* genetics by identifying high confidence hits with significantly increased or decreased cell fitness, information that promises to guide future strain engineering seeking to improve production host tolerance to harsh environmental conditions.

Another approach to resolve the effect of poorly active sgRNA in CRISPR screens is through the use of a library that only contains a small set of highly active guides targeting all required genes. Leveraging sgRNA activity profiles from the first Cas9 screen, as well the sgRNA prediction algorithm DeepGuide discussed in the previous chapter, we designed an optimized Cas9 library that targeted every protein coding gene in the genome at a 3-fold coverage. MNase-Seq was performed to determine strain specific nucleosome occupancy information, and DeepGuide made use of this epigenetic feature for guide activity predictions that resulted in the optimized library. The optimized Cas9 library was half the size of the original Cas9 library, thus easing the transformation efficiency requirements of pooled screens. This library, with only a small activity correction that removed 18% of the sgRNA, was able to capture 75% of the consensus set. Its performance was comparable to the performance of the acCRISPR predictions of the first Cas9 library, which captured nearly 85% of the consensus set, at half the library size. However, the lower performance than expected is indicative of a higher fraction of poorly performing sgRNA (Supplementary Fig. x) from predictions which we anticipate can be improved as guide activity prediction algorithms become further refined.

acCRISPR is an end-to-end pipeline for the analysis of pooled CRISPR screens. It takes a hybrid approach that combines experimental and computational methods to determine the activity of each guide in a pooled CRISPR screen and uses this information to correct screening outcomes based on guide activity. We use this pipeline to generate new knowledge on the genetics of *Y. lipolytica*, including the identification of a consensus set of essential genes for growth on glucose and for calling loss and gain of fitness hits for

growth under environmental stress conditions. The optimized library designed in this chapter is a culmination of all our previous work on CRISPR screens. This library consists of mostly highly active guides and is capable of matching the hit calling accuracy of the older Cas9 library at half its size. While this work focuses on analyzing screens conducted in *Y. lipolytica*, the same experimental-computational workflow can be readily applied to other organisms in which accurate computational prediction or genome-wide functional screens can be used to estimate sgRNA activities.

5.5 Materials and Methods

5.5.1 acCRISPR framework

acCRISPR performs essential gene identification by calculating two scores for each sgRNA, namely the *cutting score* (CS) and the *fitness score* (FS). CS and FS are the $\log_2$-fold change of sgRNA abundance in the appropriate treatment sample with respect to that in the corresponding control sample (see **Supplementary File 5.13** for replicate correlations of sgRNA abundance in control and treatment samples for Cas9 and Cas12a screens). Let us call C_i and T_i the control and treatment samples, respectively, for determining cutting scores. The cutting score CS_i of sgRNA i is defined as follows

$$CS_i = -log_2\left(\frac{\bar{x}_{T_1,i}}{\bar{x}_{C_1,i}}\right)$$

where xC1,i and xT1,i indicate the total normalized read counts of sgRNA i in samples C_i and T_i, respectively, averaged across all replicates in their respective samples. A

pseudocount of one is added to each raw count before normalization to prevent division by zero.

Similarly, let us call C_2 and T_2 control and treatment samples, respectively, for the estimation of the fitness score. The fitness score FS_i of sgRNA *i* is defined as follows

$$FS_i = log_2\left(\frac{\bar{x}_{T_2,i}}{\bar{x}_{C_2,i}}\right)$$

where xC2,i and xT2,i are average total normalized read counts in samples C_2 and T_2, respectively, for sgRNA *i*. FS_i represents the change in fitness when a gene targeted by sgRNA *i* is knocked out.

Given a CS-threshold *T*, acCRISPR creates a *CS-corrected library* by removing any sgRNA from the original library that has a cutting score less than *T*. However, if no sgRNA for a given gene has a CS that exceeds *T,* the sgRNA with the highest CS that targets that gene is kept in the CS-corrected library.

The fitness score FS_g for a gene *g* is calculated as the average of fitness scores of all sgRNA targeting gene *g*, as follows

$$FS_g = \frac{\sum_{i\epsilon g} FS_i}{m_g}$$

where mg represents the total number of sgRNA targeting gene *g* in the CS-corrected library. FS_g indicates the overall change in fitness in a particular screening condition when

gene *g* is knocked out. Since the knockout of an essential gene reduces cell fitness, essential genes would have lower fitness scores compared to non-essential genes.

acCRISPR identifies essential genes from a screening dataset by first creating a null distribution and then computing a p-value. The null distribution is assumed to be Gaussian with mean μ and standard deviation σ. This distribution represents the population of fitness scores of non-essential genes. Previous studies on essential gene identification in different yeasts have found ~20% of genes in the yeast genome to be typically essential for growth [19–21]. Thus we hypothesize that genes having FS values higher than the 20th percentile in the screening dataset are putatively non-essential. The value of μ is assumed to be equal to the median of all gene FS values and σ is computed as follows:

(i) 1000 putatively non-essential genes are randomly sampled and sgRNA targeting these genes are pooled together to form an 'sgRNA pool.'

(ii) A set of *N* sgRNA are randomly sampled from this pool and assumed to target a pseudogene, the FS of this pseudogene is calculated as the average fitness score of the sampled sgRNA. This step is repeated to generate a total of 1000 pseudogenes.

(iii) The standard deviation of the fitness scores of these 1000 pseudogenes is computed.

(iv) Steps (i)-(iii) are repeated 50 times and σ of the null distribution is calculated as the average of the 50 standard deviations (obtained in step (iii)).

(v) In these calculations, the value of *N* is initialized to the average coverage of the original library rounded off to the nearest integer. If the total number of sgRNA to be sampled from the sgRNA pool (using this value of N) is more than twice the pool size, *N* is reduced until this value drops below 2.

To identify essential genes, the resulting null distribution is used to perform a one-tailed z-test of significance for every gene in the dataset to determine whether its fitness score is significantly lower than μ. The raw p-values from the z-test are adjusted for multiple comparisons by FDR-correction and genes having corrected p-values less than a certain threshold (default: 0.05) are deemed as essential. Since every CS-threshold would result in a different essential gene set, the final set of essential genes is decided based on the value of a metric called the 'ac-coefficient', which is defined as:

$$ac - coefficient = (CS - cutoff) * (avg.coverage\ of\ the\ CS - corrected\ library)$$

The CS-threshold at which the ac-coefficient is maximum is considered optimum, and the set of essential genes obtained at this threshold is taken at the final essential gene set. In order to find the maximum ac-coefficient amongst values at different CS-thresholds, only those thresholds should be considered at which the average coverage of the library is greater than 2, since a genome coverage of less than 2 would reduce statistical power to accurately determine gene essentiality.

For analyzing stress tolerance data to identify loss- and gain-of-function hits (LOF and GOF), acCRISPR calculates a tolerance score (TS) per sgRNA and per gene in the same manner as FS. The fraction of genes directly related to stress tolerance is typically less than the number of essential genes. Thus, we assume that 95% of genes in the screening dataset (*i.e.,* TS values between the 2.5th percentile and 97.5th percentile) are putatively non-significant, and use them for calculating the null distribution parameters (μ and σ). Further, acCRISPR uses a two-tailed test of significance to identify LOF and GOF hits.

5.5.2 Implementation of acCRISPR with different input datasets

acCRISPR takes raw sgRNA counts from genome-wide screens as input and processes them to calculate CS and FS per sgRNA, as described in the previous section. However, if CS and FS values have already been calculated previously or are readily available, they can be directly provided as input by skipping log_2-fold change calculation from raw counts.

For the CRISPR-Cas9 and -Cas12a datasets, acCRISPR was first implemented using raw sgRNA counts for all targeting sgRNA in the libraries. In subsequent acCRISPR runs, CS and FS values from the first run were input to the method (*i.e.*, log_2-fold change calculation was skipped) along with a CS-threshold to identify essential genes using a CS-corrected library. For essential gene identification, a one-tailed test of significance was performed.

For implementing acCRISPR using guide activity scores from prediction algorithms, the predicted activity of each guide was provided in place of an experimentally derived CS

value along with FS as input for each run. Guide activity and CS thresholds used for analyzing datasets can be found in **Supplementary Table S5.1**.

In the tolerance datasets, raw sgRNA counts for CS calculation from CRISPR-Cas9 growth screening dataset were used in conjunction with raw counts for TS calculation from the specific screening condition. Significant genes were determined by performing a two-tailed test of significance. In all cases, genes having FDR-corrected p-value less than 0.05 were considered as significant.

5.5.3 Implementation of other CRISPR screen analysis methods

For implementing JACKS [10] and CRISPhieRmix [18], PO1f and PO1f Cas9/Cas12a strains of *Y. lipolytica* were used as control and treatment samples respectively.

Raw sgRNA counts from these two strains were provided as input to JACKS v0.2. To obtain p-values from JACKS, 500 genes classified as 'non-essential' by the transposon analysis [19] were randomly sampled and provided separately as negative control genes for the CRISPR-Cas9 and -Cas12a datasets. The raw p-values were FDR-adjusted and genes having a corrected p-value less than 0.05 were deemed as essential.

Raw sgRNA counts from untransformed library samples were used as control (initial sgRNA abundance) and those from PO1f Cas9/Cas12a were used as treatment for MAGeCK-VISPR v0.5.6 [11]. Since the data being analyzed came from LOF screens, two-tailed raw p-values from Wald test were converted to one-tailed p-values, followed by

FDR-correction. Genes having FDR-adjusted p-value less than 0.05 were considered as essential.

CRISPhieRmix v1.1 was implemented using R 4.0.2 (Rstudio 1.4.1106) by providing log_2-fold changes of all sgRNA as input. The log_2-fold changes were calculated in a manner similar to fitness scores. Log_2-fold changes of non-targeting sgRNA in the respective libraries were provided as negative controls. The parameter *screenType* was set to 'LOF' since the sgRNA log_2-fold changes were obtained from LOF screens. Genes having FDR-adjusted (1 – *genePosteriors*) values less than 0.05 were deemed as essential.

5.5.4 Microbial strains and culturing

All strains used in this work are presented in **Supplementary Table S5.2**. We describe the parent *Yarrowia* strain used for molecular cloning, and the related culture conditions here.

Yarrowia lipolytica PO1f (MatA, *leu2-270*, *ura3-302*, *xpr2-322*, *axp-2*) is the parent for all mutants used in this work. Cas9 and Cas12a expressing strains were constructed by integrating UAS1B8-TEF(136)-Cas9-CYCt and UAS1B8-TEF(136)-LbCpf1-CYCt expression cassettes into the A08 locus [9,46]. The PO1f Cas9 *ku70* and PO1f Cas12a *ku70* strains were constructed by disrupting *KU70* using CRISPR-Cas9 as previously described [17].

Yeast culturing was conducted at 30 °C in 14 mL polypropylene tubes or 250 mL baffled flasks as noted, at 225 RPM. Under non-selective conditions, *Y. lipolytica* was

grown in YPD (1% Bacto yeast extract, 2% Bacto peptone, 2% glucose). Cells transformed with sgRNA-expressing plasmids were selected for in synthetic defined 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). CRISPR screens for determining tolerance to high salinity were done in SD-leu containing a final concentration of 0.75M and 1.5M sodium chloride. The desired salinity was achieved by the addition of an appropriate quantity of autoclaved 5M sodium chloride stock solution. CRISPR screens for determining tolerance to acidity were done in SD-leu media with the pH adjusted to 3 and 2.5 using citric acid and sodium hydroxide. To attain a pH of 2.5, the SD-leu media contained a final concentration of 50mM of citric acid. To obtain a pH of 3, the media was first set to a pH of 2.5 with 50mM of citric acid and 1M sodium hydroxide was added dropwise until the desired set point was reached.

All plasmid construction and propagation were conducted in *Escherichia coli* TOP10. Cultures were conducted in Luria-Bertani (LB) broth with 100 mg L^{-1} ampicillin at 37 °C in 14 mL polypropylene tubes, at 225 RPM. Plasmids were isolated from *E. coli* cultures using the Zymo Research Plasmid Miniprep Kit.

5.5.5 Plasmid construction

All plasmids and primers used in this work are listed in **Supplementary Tables S5.3 and S5.4**. The plasmids used to construct Cas9 and Cas12a expressing strains of Y. lipolytica PO1f and the sgRNA expression plasmids were previously reported (see refs. [9]

and [16]). We describe the construction of these plasmids again here to provide a complete accounting of this work.

For *CAS9* integration, we constructed the vector pHR_A08_Cas9, which integrates a UAS1B8-Cas9 expression cassette into the A08 locus of *Y. lipolytica* PO1f. First, 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). Integration was accomplished as previously described using a two plasmid CRISPR-mediated markerless approach [46]. The creation of the Cas9 genome-wide library expression plasmid was facilitated by removing the Cas9-containing fragment from pCRISPRyl using restriction enzymes BamHI and HindIII, and circularizing. The M13 forward primer was used to ensure correct assembly of the construct.

LbCAS12a integration was accomplished in a similar manner. We first constructed pHR_A08_LbCas12a by digesting pHR_A08_hrGFP (Addgene #84615) with BssHII and NheI, and the LbCAS12a fragment was inserted using the New England BioLab (NEB) NEBuilder® HiFi DNA Assembly Master Mix. The *LbCAS12a* gene fragment was amplified along with the necessary overlaps by PCR using Cpf1-Int-F and Cpf1-Int-R primers from pLbCas12ayl. Successful cloning of the LbCas12a fragment was confirmed with sequencing primers A08-Seq-F, A08-Seq-R, Tef-Seq-F, Lb1-R, Lb2-F, Lb3-F, Lb4-F, and Lb5-F. To create the Cas12a sgRNA genome-wide library expression plasmid (pLbCas12ayl-GW) the UAS1B8-TEF- LbCas12a-CYC1 fragment was removed from pLbCas12ayl with the use of XmaI and HindIII restriction enzymes. Subsequently, the

primers BRIDGE-F and BRIDGE-R were used to circularize the vector, and the M13 forward primer was used to ensure correct assembly of the construct.

The gRNAs library vector was constructed using pCas9yl-GW (SCR1'-tRNA-AvrII site) as the backbone. The library was generated by digesting pCRISPRyl with BamHI and HindIII and circularizing to remove the Cas9 gene and its promoter and terminator using (NEBuilder® HiFi DNA Assembly). The methods used to create the guide library are provided below in the sgRNA library cloning subsection.

The LbCas12a sgRNA expression plasmid (pLbCas12ayl) was similarly constructed, but a second direct repeat sequence at the 5' of the polyT terminator in pCpf1_yl (see ref [16]) was added. This was done to ensure that library sgRNAs could end in one or more thymine residues without being constructed as part of the terminator. To make this mutation, pCpf1_yl was first linearized by digestion with SpeI. Subsequently, primers ExtraDR-F and ExtraDR-R were annealed and this double-stranded fragment was used to circularize the vector (NEBuilder® HiFi DNA Assembly).

5.5.6 sgRNA library design

sgRNA library design for the Cas9 and Cas12a CRISPR systems was accomplished as previously described in refs. [9] and [16]. The critical elements of the design are described again here.

Using the annotated genome of PO1f's parent strain (CLIB89; [https://www.ncbi.nlm.nih.gov/assembly/GCA_001761485.1] [47]) as a reference, custom

MATLAB scripts were used to design up to 8 unique Cas12a sgRNAs per gene. First, a list of all sgRNAs (25 nucleotides in length) with a TTTV (V=A/G/C) PAM were identified in both the top and bottom strand of each CDS (List A). A second list containing all possible 25nt sgRNAs with a TTTN (N=any nucleotide) PAM from the top and bottom strands of all 6 chromosomes in *Y. lipolytica* was also generated and used as a reference set to test for sgRNA uniqueness (List B). The uniqueness test was carried out by comparing the first 14nt of each sgRNA (seed sequence) in List A to the first 14nt of every sgRNA in List B. Any sequence that occurred more than once was deemed as not-unique and was removed from List A. sgRNAs that passed the uniqueness test were then picked in an unbiased manner, with even representation from the top and bottom strands when possible, starting from the 5' end of the CDS. When possible 8 unique sgRNAs were selected for each gene. In cases where 8 unique guides were not available, all unique guides were selected. In addition to the gene targeting guides, 651 non-targeting control guides were also designed. Random 25nt sequences were generated and each sequence was queried against the PO1f genome. Only sgRNA sequences in which the first 10nt were not found anywhere in the genome were selected and used as part of the control set.

The Cas9 sgRNA library was similarly designed, with the following differences. Working with the annotated CLIB89 genome, custom MATLAB scripts were used to identify unique sgRNAs (NGG PAM + 12 bp closest to the PAM) located within the first 300 bp of the gene. Subsequently, the top 6 sgRNAs from this filtered list were ranked based on their on-target activity score (Designer v1 [25]) and the top 6 guides were selected. 480 sgRNAs with random sequence were also added to the library as non-targeting

controls. These guides were confirmed not to target anywhere within the genome by ensuring that the first 12 nt of the sgRNA did not map to any genomic locus [9].

Custom MATLAB scripts were used to design the optimized Cas9 library, and the crucial elements of the design are reported here. The optimized library had 3 guides designed for all 7919 mRNA coding genes in the CLIB89 genome. Of these 3 guides, 2 were intended to be picked from the pool of best performing guides in the previous Cas9 screen, while the third guide was designed by DeepGuide predictions. First, sgRNA with CS>4.0 from the first Cas9 screen were filtered and the best two sgRNA for each gene were identified and saved, if present. The third sgRNA for all genes, as well as guides for any genes that did not have two highly active guides (CS>4.0) from the first screen were instead obtained DeepGuide's best predictions for that gene. DeepGuide predictions were enabled by nucleosome occupancy scores for all guides within all CDS in the CLIB89 genome, presented in **Supplementary File 5.16**. These nucleosome occupancy scores were derived from an MNase-Seq experiment performed in the PO1f strain. Please refer to the subsection **x**, of the methods for further details regarding the experimental methods, library preparation, and data analysis to obtain per base nucleosome occupancy scores. All sgRNA in the optimized library were verified to contain unique a seed sequence (11 nucleotides closest to the PAM). 360 nontargeting sgRNA which showed the poorest cutting scores from the prior Cas9 screen were selected as nontargeting sgRNA for this library. These guides were confirmed not to target anywhere within the genome by ensuring that the first 12 nt of the sgRNA did not map to any genomic locus.

5.5.7 sgRNA library cloning

The Cas12a library targeting the protein-coding genes in PO1f was ordered as an oligonucleotide pool from Agilent Technologies Inc. and cloned in-house using the Agilent SureVector CRISPR Library Cloning Kit (Part Number G7556A) as previously described in [16].

First, the backbone pLbCas12ayl-GW was linearized and amplified by PCR using the primers InversePCR-F and InversePCR-R. To verify the completely linearized vector, we DpnI digested amplicon, purified the product with Beckman AMPure XP SPRI beads, and transformed it into *E. coli* TOP10 cells. A lack of colonies indicated a lack of contamination from the intact backbone.

Library ssDNA oligos were then amplified by PCR using the primers OLS-F and OLS-R for 15 cycles as per vendor instructions using Q5 high fidelity polymerase. The amplicons were cleaned using the AMPure XP beads prior to use in the following step. sgRNA library cloning was conducted in four replicate tubes using Agilent's SureVector CRISPR library cloning kit (Catalog #G7556A). The completed reactions were pooled and subjected to another round of cleaning.

Two amplification bottles containing 1L of LB media and 3 g of high-grade low-gelling agarose were prepared, autoclaved, and cooled to 37 °C (Agilent, Catalog #5190-9527). Eighteen replicate transformations of the cloned library were conducted using Agilent's ElectroTen-Blue cells (Catalog #200159) via electroporation (0.2 cm cuvette, 2.5 kV, 1 pulse). Cells were recovered and with a 1 hr outgrowth in SOC media at 37 °C (2%

tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, and 20 mM glucose.) The transformed *E. coli* cells were then inoculated into two amplification bottles and grown for two days until colonies were visible in the matrix. Colonies were recovered by centrifugation and subject to a second amplification step by inoculating an 800 mL LB culture. After 4 hr, the cells were collected, and the pooled plasmid library was isolated using the ZymoPURE II Plasmid Gigaprep Kit (Catalog #D4202) yielding ~2.4 mg of plasmid DNA encoding the Cas12a sgRNA library. The library was subject to a NextSeq run to test for fold coverage of individual sgRNA and skew.

The Cas9 library was constructed by the US Department of Energy's Joint Genome Institute as a deliverable of Community Science Project (CSP) 503076. Experimental details as previously described in ref [9] are included here for completeness. The pooled sgRNA library targeting the protein-coding genes of PO1f was ordered as four oligo pools each consisting of 25% of the designed sgRNAs from Twist Bioscience and cloned. The separation into different sub-libraries was done to test different methods of assembly; the details of each approach are briefly described here.

For sub-libraries 1 and 3, second-strand synthesis reactions were conducted using the primer sgRNA-Rev2 and T4 DNA polymerase (NEB), gel extracted, and purified using Zymo Research Zymo-Spin 1 columns. For sub-libraries 2 and 4, oligos were amplified with primers via Q5 DNA polymerase (NEB) using 0.2 picomoles of DNA as a template for 7 cycles, and column purified. Library 2 had overlaps of 20 bp on either side of the

spacer and was amplified with 60mer_pool-F and spacer-AarI.rev. Library 4 had overlaps of ~60 bp on either side of the spacer and was amplified with primers pLeu-mock-sgRNA.fwd and sgRNA-Rev2. Libraries 1, 3, and 4 were cloned into the AarI digested pCas9yl-GW vector using the Gibson Assembly HiFi HC 1-step Master Mix (SGI-DNA). Library 2 was digested with AarI and cloned into pCas9yl-GW digested with AarI using Golden Gate assembly with T4 DNA ligase (NEB).

The cloning method for library 4 resulted in the least number of spacers missing in the propagated library. 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 in colonies for the number of library variants. The plasmid library was isolated from the transformed cells after a short outgrowth.

The optimized Cas9 library was cloned in a manner similar to the Cas12a library by making use of the Agilent SureVector CRISPR Library Cloning Kit (Part Number G7556A). Briefly, the backbone pCas9yl-GW was linearized and amplified by PCR using the primers InversePCRCas9Opt-F and InversePCRCas9Opt-R. To verify the completely linearized vector, we DpnI digested amplicon, purified the product with Beckman AMPure XP SPRI beads, and transformed it into *E. coli* TOP10 cells. A lack of colonies indicated a lack of contamination from the intact backbone. Library ssDNA oligos were then amplified by PCR using the primers OLS-F and OLS-R for 15 cycles as per vendor instructions using Q5 high fidelity polymerase. The amplicons were cleaned using the AMPure XP beads prior to use in the following step. sgRNA library cloning was conducted

in four replicate tubes and subsequently, pooled and cleaned up as per manufacturer's instructions.

One amplification bottle containing 1L of LB media and 3 g of high-grade low-gelling agarose was prepared, autoclaved, and cooled to 37 °C (Agilent, Catalog #5190-9527). Ten transformations of the cloned library were conducted using Agilent's ElectroTen-Blue cells (Catalog #200159) via electroporation (0.2 cm cuvette, 2.5 kV, 1 pulse). Cells were recovered and with a 1 hr outgrowth in SOC media at 37 °C (2% tryptone, 0.5% yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM $MgCl_2$, 10 mM $MgSO_4$, and 20 mM glucose.) The transformed *E. coli* cells were then inoculated into the amplification bottle and grown for two days until colonies were visible in the matrix. Colonies were recovered by centrifugation and subject to a second amplification step by inoculating two 250 mL LB cultures. After 4 hr, the cells were collected, and the pooled plasmid library was isolated using the ZymoPURE II Plasmid Gigaprep Kit (Catalog #D4202) yielding ~1.8 mg of plasmid DNA encoding the Cas12a sgRNA library. The library was subject to a NextSeq run to test for fold coverage of individual sgRNA and skew.

5.5.8 Yeast transformation and screening

Transformation of the Cas9 and Cas12a sgRNA plasmid libraries into *Y. lipolytica* was done using a method previously described in refs. [9,16]. For Cas12a experiments, 3 mL of YPD was inoculated with a single colony of the strain of interest and grown in a 14 mL tube at 30 °C with shaking at 200 RPM for 22-24 hours (final OD ~30). Cells were pelleted by centrifugation (6,300g), washed with 1.2 mL of transformation buffer (0.1 M LiAc, 10

mM Tris (pH=8.0), 1 mM EDTA), pelleted again by centrifugation, and resuspended in 1.2 mL of transformation buffer. To these resuspended cells, 36 µL of ssDNA mix (8 mg/mL Salmon Sperm DNA, 10 mM Tris (pH=8.0), 1 mM EDTA), 180 µL of β-mercaptoethanol mix (5% β-mercaptoethanol, 95% triacetin), and 8 µg of plasmid library DNA were added, mixed via pipetting, and incubated for 30 mins. at room temperature. After incubation, 1800 µL of PEG mix (70% w/v PEG (3,350 MW)) was added and mixed via pipetting, and the mixture was incubated at room temperature for an additional 30 min. Cells were then heat shocked for 25 min at 37 °C, washed with 25 mL of sterile Milli-Q H_2O, and used to inoculate 50 mL of SD-leu media. Dilutions of the transformation (0.01% and 0.001%) were plated on solid SD-leu media to calculate transformation efficiency. Three biological replicates of each transformation were performed for each condition. Transformation efficiency for each replicate from the Cas9 and Cas12a experiments is presented in **Supplementary Table S5.5**.

Transformation for the first Cas9 library as well as the optimized Cas9 library was done in a very similar manner. Briefly, half the amount of cells, DNA, and other chemical reagents described above were used for a single transformation and multiple transformations were done and pooled as necessary to ensure adequate diversity to maintain library representation and minimize the effect of plasmid instability (100x coverage, 5 x 10^6 total transformants per biological replicate).

Screening experiments were conducted in 25 mL of liquid media in a 250 mL baffled flask (220 RPMshaking, 30 °C). Cells first reached confluency after two days of growth

(OD_{600} ~12), at which time 200 µL, which includes a sufficient number of cells for approximately 500-fold library coverage, was used to inoculate 25 mL of fresh media. The cells were again subcultured upon reaching confluency after four days of culture, and the experiment was stopped after reaching confluency again on day six of the screen. Glycerol stocks of day 2 cultures were also prepared and used to start other growth screens as discussed in a following subsection.

On days two, four, and six, 1 mL of culture was removed to isolate sgRNA expression plasmids for deep sequencing. Each sample was first treated with DNase I (New England Biolabs; 2 µL and 25µL of DNaseI buffer) for 1 h at 30 °C to remove any extracellular plasmid DNA. Cells were then isolated by centrifugation at 4,500g, and the resulting cell pellets were stored at -80 °C prior to sequencing.

5.5.9 *Y. lipolytica* pH and salt tolerance screens

CRISPR-Cas9 growth screens with high salinity and low pH were conducted in synthetic defined media deficient in leucine. Media were prepared with two different salt and citric acid concentrations as defined in the microbial strains and culturing subsection. 150 uL (approximately $1x10^7$ cells) of Day 2 glycerol stocks of PO1f Cas9 strain transformed with the sgRNA library were used to inoculate 250 mL baffled flasks containing 25 mL of five different media: SD-leu, SD-leu (0.75M NaCl), SD-leu (1M NaCl), SD-leu (pH 2.5) and SD-leu (pH 3). Three biological replicates were cultured for each different media condition. Outgrowth following inoculation was done at 30 °C at 225 RPM. Cells were grown for two days, and fresh media was inoculated with at least $1x10^7$

cells and grown for another two days. The experiment was halted after 4 days of outgrowth following inoculation. On the last day, 1 mL of culture was removed, treated with DNase I, pelleted, and processed to extract plasmids as described above. Extracted plasmids were quantified by qPCR and amplified with forward (Cr1665-Cr1668) and reverse primers (Cr1669-Cr1671, Cr1673, and Cr1709) containing the necessary barcodes and adapters for NGS using NextSeq. Growth of the PO1f Cas9 strain in SD-leu was used as a control in the tolerance screens to select for genetic perturbations that either conferred a growth advantage or disadvantage only under the stressed condition.

5.5.10 Library isolation and sequencing

Frozen culture samples from pooled CRISPR screens were thawed and resuspended in 400 µL sterile, Milli-Q H_2O. Each cell suspension was split into two, 200 µL samples. Plasmids were isolated from each sample using a Zymo Yeast Plasmid Miniprep Kit (Zymo Research). Splitting into separate samples here was done to accommodate the capacity of the Yeast Miniprep Kit, specifically to ensure complete lysis of cells using Zymolyase and lysis buffer. This step is critical in ensuring sufficient plasmid recovery and library coverage for downstream sequencing. The split samples from a single pellet were pooled, and the plasmid copy number was quantified using quantitative PCR with qPCR-GW-F and qPCR-GW-R and SsoAdvanced Universal SYBR Green Supermix (Biorad). Each pooled sample was confirmed to contain at least 10^7 plasmids so that sufficient coverage of the sgRNA library is ensured.

To prepare samples from the Cas12a screen for next-generation sequencing, isolated plasmids were subjected to PCR using forward (ILU1-F, ILU2-F, ILU3-F, ILU4-F) and reverse primers (ILU(1-12)-R) containing all necessary barcodes and adapters for next-generation sequencing using the Illumina platform (**Supplementary Table 6**). Schematics of the amplicons from the Cas9 and Cas12a screens submitted for NGS are depicted in **Supplementary Fig. S5.10**. At least 0.2 ng of plasmids (approximately $3x10^7$ plasmid molecules) were used as template for PCR and amplified for 16 cycles and not allowed to proceed to completion to avoid amplification bias. PCR product was purified using SPRI beads and tested on the bioanalyzer to ensure the correct length.

Samples from the Cas9 screens with both the old and the optimized libraries were prepared as previously described in ref [9]. Briefly, isolated plasmids were amplified using forward (Cr1665-Cr1668) and reverse primers (Cr1669-Cr1673; Cr1709-1711) containing the necessary barcodes, pseudo-barcodes, and adapters (**Supplementary Table S5.7**). Approximately $1x10^7$ plasmids were used as a template and amplified for 22 cycles, not allowing the reaction to proceed to completion. Amplicons at 250 bp were then gel extracted and tested on the bioanalyzer to ensure correct length. Samples were pooled in equimolar amounts and submitted for sequencing on a NextSeq 500 at the UCR IIGB core facility.

5.5.11 Generating sgRNA read counts from raw reads

Next-generation sequencing raw fastq files were processed using the Galaxy platform [48]. Read quality was assessed using FastQC v0.11.8., demultiplexed using Cutadapt

v1.16.6, and truncated to only contain the sgRNA using Trimmomatic v0.38. Custom MATLAB scripts were written to determine counts for each sgRNA in the library using Bowtie alignment (Bowtie2 v2.4.2; inexact matching) and naïve exact matching (NEM). The final count for each sgRNA was taken as the maximum of the two methods. A large majority of data points were derived from inexact matching with Bowtie, in only a few cases where Bowtie failed to give proper alignment, was the exact matching value used. Parameters used for each of the tools used on Galaxy for Cas12a and both Cas9 screens are provided in **Supplementary Tables S5.8 and S5.9 respectively**. MATLAB scripts are provided as part of the GitHub link found below in the "Data and software availability" section. **Supplementary File 5.14** provides further information correlating the NCBI SRA file names to the information needed for demultiplexing the readsets. Analysis of raw (unoptimized) Cas9 and Cas12a libraries revealed 721 and 12 sgRNA, respectively, that were found to be either missing or having very low normalized abundance (< 5% of the normalized mean abundance of the library) and were discarded from further analysis (see **Supplementary File 5.15** for raw sgRNA counts of the untransformed Cas9 and Cas12a libraries). Analysis of the optimized Cas9 library revealed 3 sgRNA that were either missing or having very low normalized abundance (< 5% of the normalized mean abundance of the library), indicating excellent cloning and coverage.

5.5.11 Nucleosome occupancy determination

Nucleosome occupancy in the PO1f strain of *Y. lipolytica* was determined by performing MNase-Seq. In this section, the protocol for nucleosome extraction, library preparation for sequencing on the Illumina MiSeq platform, and the bioinformatics analysis

of NGS reads to arrive at per base nucleosome occupancy scores are presented. This protocol was adapted from Methods in Enzymology chapter for nucleosome extraction in yeast [51].

Overnight cultures of PO1f in 2 mL YPD were used to inoculate larger 40 mL YPD cultures in shake flasks. When the cells reached mid exponential phase, after 11 hours of growth (OD ~3), 2.2 mL of 37% formaldehyde was added to the culture for a 2% v/v final concentration, in order to crosslink the nucleosomes to the DNA to maintain their positions in subsequent steps of the protocol. The crosslinking reaction was allowed to proceed for 15 minutes before quenching with 3 mL of 2M glycine (for a final concentration of 0.125M) to stop the crosslinking reaction. 25 OD of cells (~8 mL of culture) were collected, pelleted and taken forward for spheroplasting and nuclei extraction. Cell pellets were resuspended in 600 uL of Y-Lysis Buffer (Cat. No. Y1002-1-6), and 30 uL of Zymolyase (Cat. No. E1004) was added mixture. The tube was incubated at 37 C for 60 minutes to complete the spheroplasting reaction. A small 5 uL aliquot was taken, and an equal volume of 2.5% SDS was added. Efficient spheroplasting is indicated by the sample turning from cloudy to clear, demonstrating that the cell wall was lysed by the SDS. The tubes containing the spheroplasted cells were pelleted by centrifugation to 3000 x g for 5 min at 4 C. The supernatant was discarded, and the pellet was resuspended with 500 uL of 1.2M sorbitol. The spheroplasts were pelleted once more as previously described, supernatant discarded, and the pellet resuspended in 500 uL of Nuclei Prep Buffer (Cat. No. D5220-2). The nuclei were pelleted by centrifugation to 3000 x g for 5 min at 4 C, and the supernatant discarded.

The nuclei pellet was then resuspended in 100 uL of TakaraBio's 10X Micrococcal Nuclease buffer, and 2 uL of Micrococcal Nuclease (20U/uL) (Cat. No. 2910A) was added. The reaction was incubated at 37 C was 30 min to allow the MNase to degrade DNA unprotected by the crosslinked nucleosomes. MNase reaction was stopped by the addition of 20 uL of 5X MN stop buffer (Cat. No. D5220-4) and vortexing the mixture briefly. The DNA was then uncrosslinked from the proteins by the addition of 4.8 uL of 5M NaCl (final concentration of 0.2M) and 2 uL of Proteinase K. This reaction was incubated at 65 C overnight, and then purified using Zymo's EZ Nucleosomal DNA Prep kit. The DNA was cleanup was performed as per manufacturer's instructions and eluted in 20 uL of nuclease free water. The resulting pure DNA was then run on Agarose gels with a 100 bp ladder to check the efficiency of MNase treatment, and the resulting quality of nucleosomal DNA. Upon titration with varying amounts of MNase we observed that 40U of MNase for 25 OD of cells initially taken from culture, resulted largely in mono-nucleosomal DNA (~147 bp) with small bands of di-nucleosomal DNA (~300 bp) (**Supplementary Figure S11**). The protocol described above was performed in triplicate and the extracted DNA after quality inspection, was taken for library preparation for Illumina MiSeq sequencing.

Library preparation for Illumina MiSeq sequencing was performed using the NEBNext Ultra™ II DNA Library Prep Kit for Illumina (Cat. No. D5220-2). Illumina TruSeq adaptors were ligated to the extracted nucleosome DNA samples as per kit instructions. Upon adaptor ligation and cleanup, the three replicate samples were dual indexed with unique pairs of barcodes from NEBNext® Multiplex Oligos for Illumina® (Dual Index Primers Set 1) (Cat. No. E7600S), amplified using PCR and cleaned up using paramagnetic

beads. These samples were then evaluated for library quality on the bioanalyzer. All samples showed the expected band size of ~250 bp (150 bp mono-nucleosomal DNA, as well as an additional 100 bp of adapters and indices on either end). The samples were submitted to Genewiz for a 2x150 bp paired end sequencing run on the MiSeq.

NGS paired end fastq files from the MNase-Seq experiment were processed on Galaxy to arrive at per base nucleosome occupancy scores for the entire genome. Read quality was assessed using FastQC v0.11.8., and Trimmomatic (Galaxy Version 0.36.5) was used trim regions with quality scores of less than 20, as well as reads that were shorter than 36 bp. Bowtie (Galaxy Version 2.3.4.3) was used to align the reads to the genome using the default 'very sensitive, end-to-end' mode. One additional step in many NGS pipelines is PCR duplicate removal, where PCR duplicates arise from multiple PCR products from the same template molecule binding on the flowcell. These are often removed because there is concern that they may lead to false positive calls. Thus, Picard (Galaxy Version 2.18.2.2) was used on the output BAM files to mark and remove PCR duplicates. The BAM files were then filtered to keep only mapped reads (BAM Filter; Galaxy Version 2.4.1), and then sequence coverage for every nucleotide in the genome was generated using bamCoverage (Galaxy Version 3.3.2.0.0). Custom MATLAB scripts were then used to identify the position of every sgRNA from all protein coding CDS in the genome along with the necessary 40 nt context (10 nt context + 20 nt spacer + 3 nt NGG PAM + 7 nt context) required as input for DeepGuide activity predictions, and nucleosome occupancies for regions were calculated as the average nucleosome occupancy over the 40 nt window in

question. This was then min-max normalized to so that all nucleosome occupancy values ranged from 0 to 1. This data is reported in **Supplementary File 5.16.**

5.5.12 Gene ontology enrichment analysis

GO annotations for the CLIB89 reference genome of *Y. lipolytica* [49] were obtained from MycoCosm (mycocosm.jgi.doe.gov). GO analysis for the essential gene sets was performed using the Galaxy platform [48]. First, GO-slim annotations for CLIB89 were obtained using GOSlimmer v1.0.1. Next, the GO annotation and GO-slim annotation files were used to perform GO enrichment and GO-slim enrichment analyses respectively, using GOEnrichment v2.0.1. For this analysis, the list of essential genes from a particular dataset was provided as the study set, and the list of all genes covered by the corresponding library was provided as the population set. GO terms/GO-slim terms having FDR-corrected p-value less than 0.05 from the hypergeometric test were considered to be over-represented.

5.5.13 Finding essential gene homologs in *S. cerevisiae* and *S. pombe*

Sequences of essential genes in the *Y. lipolytica* consensus set from the CLIB89 strain were aligned to genes in *S. cerevisiae* and *S. pombe* using BLASTP. *S. cerevisiae* essential genes (phenotype:inviable) were retrieved from the Saccharomyces Genome Database (SGD), and *S. pombe* essential genes were taken from Kim et al., 2010 [21]. Pairs of query and subject sequences having > 40% identity from BLASTP were deemed as homologs.

5.5.14 Implementation of sgRNA activity prediction tools

DeepGuide predicted CS values for CRISPR-Cas9 and -Cas12a datasets were obtained using DeepGuide v1.0.0 [16]. sgRNA activity prediction scores from Designer v1 [25], Designer v2 [26], CRISPRspec [29], CRISPRscan [28], SSC [27], and uCRISPR [24] were obtained using CHOPCHOP v3 [50]. Similarly, DeepCpf1 scores were obtained using DeepCpf1 [30].

5.5.15 Calculation of sensitivity and precision

Sensitivity measures the fraction of the consensus set of essential genes that is covered by predicted essential genes from a given method and is computed as:

$$\% \text{ Sensitivity} = \text{No. of predicted essential genes overlapping with the consensus set} / \text{Size of the consensus set} * 100$$

Precision measures the fraction of predicted essential genes from a given method that overlap with the consensus set and is calculated as:

$$\% \text{ Precision} = \text{No. of predicted essential genes overlapping with the consensus set} / \text{Total no. of predicted essential genes} * 100$$

5.6 Data availability

The sgRNA sequencing data for all CRISPR-Cas9 and -Cas12a screens generated for this study have been deposited in the NCBI SRA database under accession code PRJNA857832. The sgRNA raw counts, cutting scores, and fitness scores generated in this study are provided as separate Supplementary Information and Source Data files.

5.7 Code availability

Source code for acCRISPR can be found at https://github.com/ianwheeldon/acCRISPR. This GitHub page includes system requirements, instructions for installation, and usage examples. Custom Matlab scripts that were used for the design of the Cas12a CRISPR library and processing of Illumina reads to generate sgRNA abundance for both Cas9 and Cas12a screens can also be found at the same link.

5.8 Author contributions

AR, VT, and IW conceived the idea, planned the experiments, and analyzed the data. AR, CS, and ML conducted the CRISPR-Cas9 growth and stress tolerance screens. AR conducted the CRISPR-Cas12a screens. VT analyzed all screens using acCRISPR and other essential gene methods, as well as performed GOslim-enrichment analysis for essential gene sets. VT, AT, AM, and SL predicted the activity of CRISPR-Cas9 and -Cas12a guides and analyzed the prediction data using acCRISPR. All authors wrote and edited the manuscript.

5.9 Acknowledgments

This work was supported by DOE DE-SC0019093, DOE Joint Genome Institute grant CSP-503076, NSF 1706545, NSF1803630, and NSF Plants-3D 1922642.

5.10 References

1. Lian, J., Schultz, C., Cao, M., HamediRad, M. & Zhao, H. Multi-functional genome-wide CRISPR system for high throughput genotype-phenotype mapping. *Nat. Commun.* 10, 5794 (2019).

2. Peters, J. M. *et al.* A Comprehensive, CRISPR-based Functional Analysis of Essential Genes in Bacteria. *Cell* 165, 1493–1506 (2016).

3. Sidik, S. M. *et al.* A Genome-wide CRISPR Screen in Toxoplasma Identifies Essential Apicomplexan Genes. *Cell* 166, 1423–1435.e12 (2016).

4. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* 159, 647–661 (2014).

5. Ramesh, A., Ong, T., Garcia, J. A., Adams, J. & Wheeldon, I. Guide RNA Engineering Enables Dual Purpose CRISPR-Cpf1 for Simultaneous Gene Editing and Gene Regulation in. *ACS Synth. Biol.* 9, 967–971 (2020).

6. Jensen, K. T. *et al.* Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. *FEBS Lett.* 591, 1892–1901 (2017).

7. Strohkendl, I. *et al.* Inhibition of CRISPR-Cas12a DNA targeting by nucleosomes and chromatin. *Sci Adv* 7, (2021).

8. Moreb, E. A. & Lynch, M. D. Genome dependent Cas9/gRNA search time underlies sequence dependent gRNA activity. *Nat. Commun.* 12, 5034 (2021).

9. Schwartz, C. *et al.* Validating genome-wide CRISPR-Cas9 function improves screening in the oleaginous yeast Yarrowia lipolytica. *Metab. Eng.* 55, 102–110 (2019).

10. Allen, F. *et al.* JACKS: joint analysis of CRISPR/Cas9 knockout screens. *Genome Res.* 29, 464–471 (2019).

11. Li, W. *et al.* Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biol.* 16, 281 (2015).

12. Löbs, A.-K., Schwartz, C. & Wheeldon, I. Genome and metabolic engineering in non-conventional yeasts: Current advances and applications. *Synth Syst Biotechnol* 2, 198–207 (2017).

13. Qiao, K., Wasylenko, T. M., Zhou, K., Xu, P. & Stephanopoulos, G. Lipid production in Yarrowia lipolytica is maximized by engineering cytosolic redox metabolism. *Nat. Biotechnol.* 35, 173–177 (2017).

14. Xue, Z. *et al.* Production of omega-3 eicosapentaenoic acid by metabolic engineering of Yarrowia lipolytica. *Nat. Biotechnol.* 31, 734–740 (2013).

15. Park, Y.-K., Ledesma-Amaro, R. & Nicaud, J.-M. Biosynthesis of Odd-Chain Fatty Acids in Enabled by Modular Pathway Engineering. *Front Bioeng Biotechnol* 7, 484 (2019).

16. Baisya, D., Ramesh, A., Schwartz, C., Lonardi, S. & Wheeldon, I. Genome-wide functional screens enable the prediction of high activity CRISPR-Cas9 and -Cas12a guides in Yarrowia lipolytica. *Nat. Commun.* 13, 922 (2022).

17. Schwartz, C., Frogue, K., Ramesh, A., Misa, J. & Wheeldon, I. CRISPRi repression of nonhomologous end-joining for enhanced genome engineering via homologous recombination in Yarrowia lipolytica. *Biotechnol. Bioeng.* 114, 2896–2906 (2017).

18. Daley, T. P. *et al.* CRISPhieRmix: a hierarchical mixture model for CRISPR pooled screens. *Genome Biol.* 19, 159 (2018).

19. Patterson, K. *et al.* Functional genomics for the oleaginous yeast Yarrowia lipolytica. *Metab. Eng.* 48, 184–196 (2018).

20. Cherry, J. M. The Saccharomyces Genome Database: Advanced Searching Methods and Data Mining. *Cold Spring Harb. Protoc.* 2015, db.prot088906 (2015).

21. Kim, D.-U. *et al.* Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nat. Biotechnol.* 28, 617–623 (2010).

22. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29 (2000).

23. Consortium, G. O. & Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* vol. 32 258D–261 Preprint at https://doi.org/10.1093/nar/gkh036 (2004).

24. Zhang, D., Hurst, T., Duan, D. & Chen, S.-J. Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design. *Proc. Natl. Acad. Sci. U. S. A.* 116, 8693–8698 (2019).

25. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nature Biotechnology* vol. 32 1262–1267 Preprint at https://doi.org/10.1038/nbt.3026 (2014).

26. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* 34, 184–191 (2016).

27. Xu, H. *et al.* Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* 25, 1147–1157 (2015).

28. Moreno-Mateos, M. A. *et al.* CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* 12, 982–988 (2015).

29. Alkan, F., Wenzel, A., Anthon, C., Havgaard, J. H. & Gorodkin, J. CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biol.* 19, 177 (2018).

30. Kim, H. K. *et al.* Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. *Nat. Biotechnol.* 36, 239–241 (2018).

31. Thorwall, S., Schwartz, C., Chartron, J. W. & Wheeldon, I. Stress-tolerant non-conventional microbes enable next-generation chemical biosynthesis. *Nat. Chem. Biol.* 16, 113–121 (2020).

32. Zhang, S., Jagtap, S. S., Deewan, A. & Rao, C. V. pH selectively regulates citric acid and lipid production in Yarrowia lipolytica W29 during nitrogen-limited growth on glucose. *J. Biotechnol.* 290, 10–15 (2019).

33. Adler, L., Blomberg, A. & Nilsson, A. Glycerol metabolism and osmoregulation in the salt-tolerant yeast Debaryomyces hansenii. *J. Bacteriol.* 162, 300–306 (1985).

34. Bahieldin, A. *et al.* Control of glycerol biosynthesis under high salt stress in Arabidopsis. *Funct. Plant Biol.* 41, 87–95 (2013).

35. Chang, Y.-L. *et al.* Yeast Cip1 is activated by environmental stress to inhibit Cdk1-G1 cyclins via Mcm1 and Msn2/4. *Nat. Commun.* 8, 56 (2017).

36. Tkach, J. M. *et al.* Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. *Nat. Cell Biol.* 14, 966–976 (2012).

37. Tokuoka, K. Sugar- and salt-tolerant yeasts. *J. Appl. Bacteriol.* 74, 101–110 (1993).

38. Espinoza, C., Liang, Y. & Stacey, G. Chitin receptor CERK1 links salt stress and chitin-triggered innate immunity in Arabidopsis. *Plant J.* 89, 984–995 (2017).

39. Gigli-Bisceglia, N. & Testerink, C. Fighting salt or enemies: shared perception and signaling strategies. *Curr. Opin. Plant Biol.* 64, 102120 (2021).

40. Dudley, A. M., Janse, D. M., Tanay, A., Shamir, R. & Church, G. M. A global view of pleiotropy and phenotypically derived gene function in yeast. *Mol. Syst. Biol.* 1, 2005.0001 (2005).

41. Park, S. G., Cha, M. K., Jeong, W. & Kim, I. H. Distinct physiological functions of thiol peroxidase isoenzymes in Saccharomyces cerevisiae. *J. Biol. Chem.* 275, 5723–5732 (2000).

42. Imkeller, K., Ambrosi, G., Boutros, M. & Huber, W. gscreend: modelling asymmetric count ratios in CRISPR screens to decrease experiment size and improve phenotype detection. *Genome Biol.* 21, 53 (2020).

43. Moreb, E. A. & Lynch, M. D. A Meta-Analysis of gRNA Library Screens Enables an Improved Understanding of the Impact of gRNA Folding and Structural Stability on CRISPR-Cas9 Activity. *CRISPR J* 5, 146–154 (2022).

44. Chao, M. C., Abel, S., Davis, B. M. & Waldor, M. K. The design and analysis of transposon insertion sequencing experiments. *Nat. Rev. Microbiol.* 14, 119–128 (2016).

45. Gale, A. N. *et al.* Identification of Essential Genes and Fluconazole Susceptibility Genes in by Profiling Transposon Insertions. *G3* 10, 3859–3870 (2020).

46. Schwartz, C., Shabbir-Hussain, M., Frogue, K., Blenner, M. & Wheeldon, I. Standardized Markerless Gene Integration for Pathway Engineering in Yarrowia lipolytica. *ACS Synth. Biol.* 6, 402–409 (2017).

47. Magnan, C. *et al.* Sequence Assembly of Yarrowia lipolytica Strain W29/CLIB89 Shows Transposable Element Diversity. *PLoS One* 11, e0162363 (2016).

48. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544 (2018).

49. Grigoriev, I. V. *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* 42, D699–704 (2014).

50. Labun, K. *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* 47, W171–W174 (2019).

51. Rando, Oliver J. "Genome-wide mapping of nucleosomes in yeast." Methods in enzymology. Vol. 470. Academic Press, 2010. 105-118.

5.11 Supplementary Information

5.11.1 Supplementary Figures

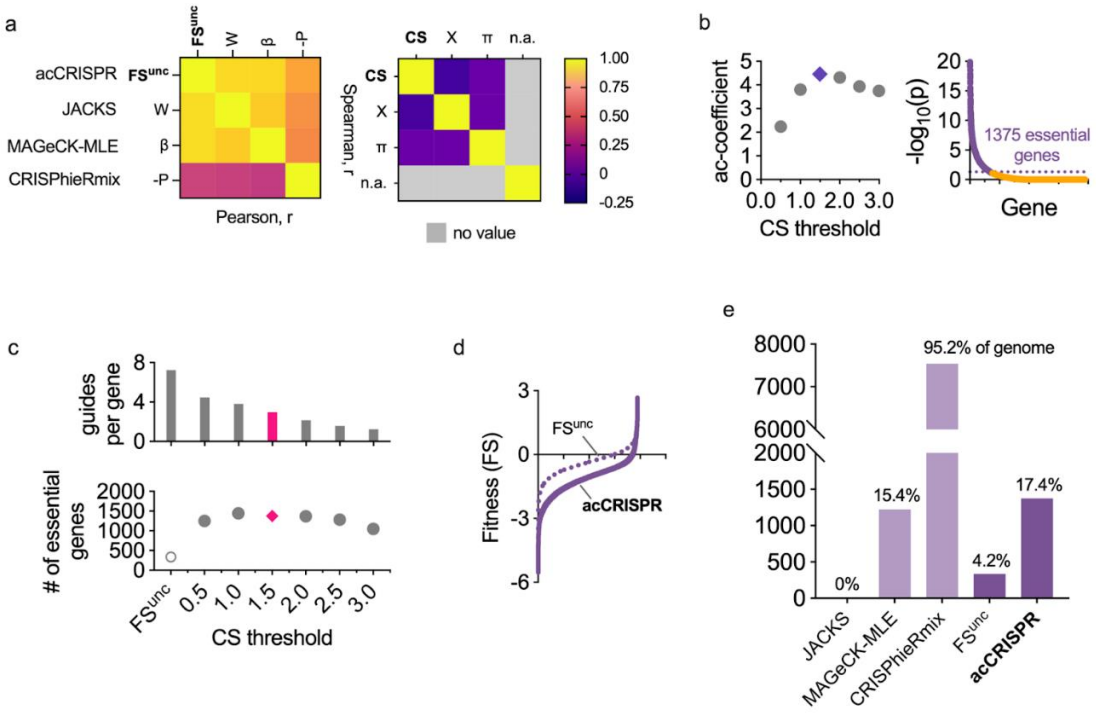

Figure S5.1. acCRISPR analysis of Cas12a growth screens in *Yarrowia lipolytica*. (**a**) Heat-maps showing Pearson (below diagonal) and Spearman (above diagonal) coefficients of fitness effects (uncorrected FS (FS^{unc}), W, β & -P; left) and sgRNA cutting efficiencies (CS, X and π; right) and from acCRISPR and three established essential gene identification algorithms, JACKS, MAGeCK-MLE and CRISPhieRmix. **(b)** ac-coefficient is calculated with increasing CS threshold values and maximum value is represented by the purple datapoint. Genes with a p-value < 0.05 were classified as essential at the maximum ac-coefficient value. (**c**) Average number of sgRNA per gene and the number of essential genes predicted with increasing CS threshold. The number of essential genes predicted for the corrected and uncorrected analyses. The data points colored in pink are the guides per gene and number of essential genes determined at the optimum CS threshold. (**d**) Fitness scores of genes with (solid line) and without (dashed line) acCRISPR processing with a CS threshold of 1.5. (**e**) Number of essential genes identified by JACKS [1], MAGeCK-MLE [2], CRISPhieRmix [3], FS^{unc}, and acCRISPR along with the percentage of total genes in the genome are reported.

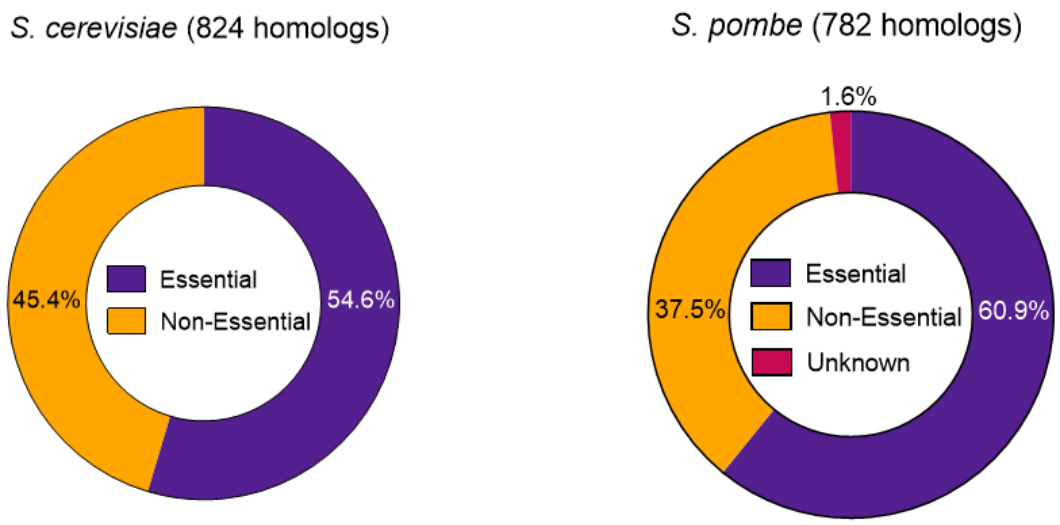

Figure S5.2. Essential gene comparison to *S. cerevisiae* and *S. pombe*. Pie charts indicating the percentage of homologs in the *Y. lipolytica* consensus set that are essential, non-essential and have unknown essentiality in *S. cerevisiae* (824 homologs) and *S. pombe* (782 homologs).

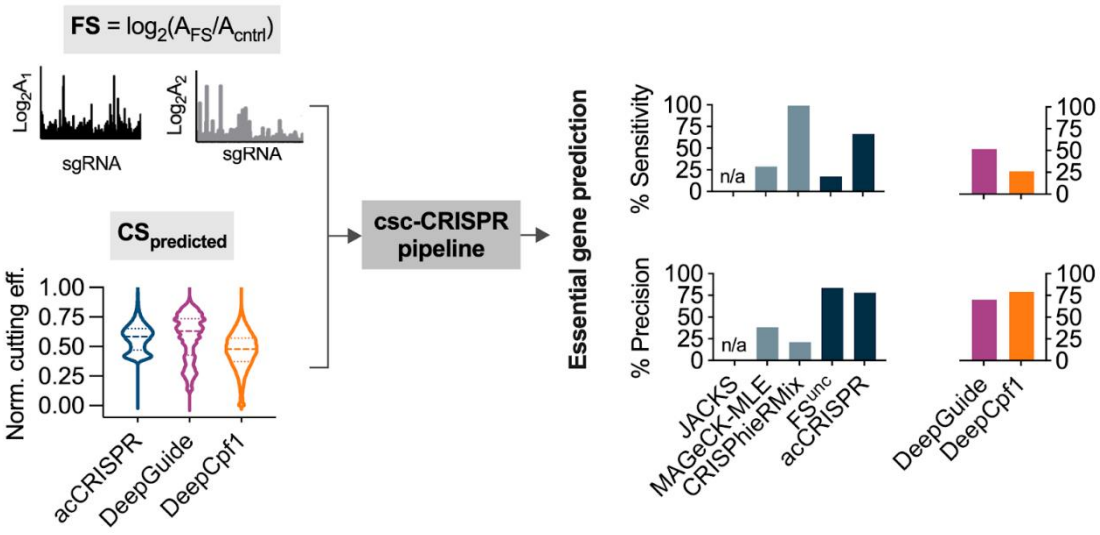

Figure S5.3. Performance of acCRISPR on the Cas12a screening dataset with predicted sgRNA activities. Essential genes were determined with acCRISPR utilizing FS along with predicted sgRNA activities from DeepGuide [4] and DeepCpf1 [5]. The violin plot shows min-max normalized sgRNA activity distributions of experimental CS determined by acCRISPR and those from DeepGuide and DeepCpf1. The % sensitivity and % precision in identifying genes from the consensus set is shown (right). Bars indicate the values of these two metrics for each prediction tool as well as for JACKS [1], MAGeCK-MLE [2], CRISPhieRmix [3], uncorrected FS (FS only) and acCRISPR.

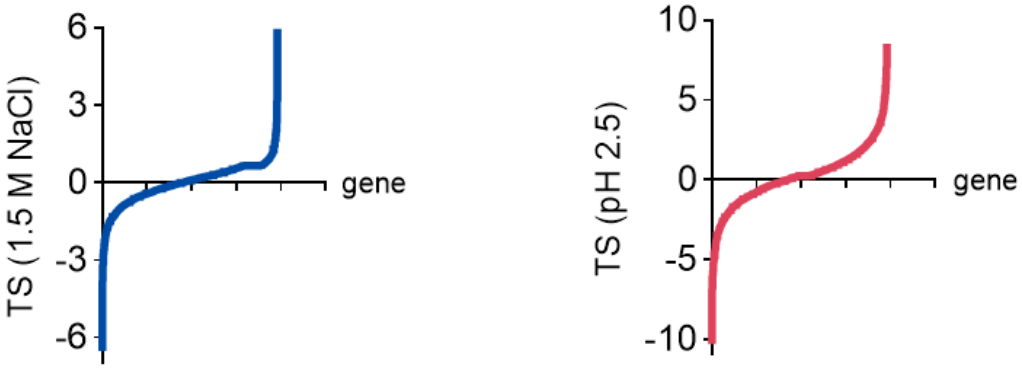

Figure S5.4. acCRISPR corrected Tolerance Scores (TS) for 1.5 M NaCl and pH 2.5 tolerance screens. S-curves showing tolerance scores of genes at a CS threshold of 4.5 for two stress conditions - 1.5 M NaCl (left) and pH 2.5 (right).

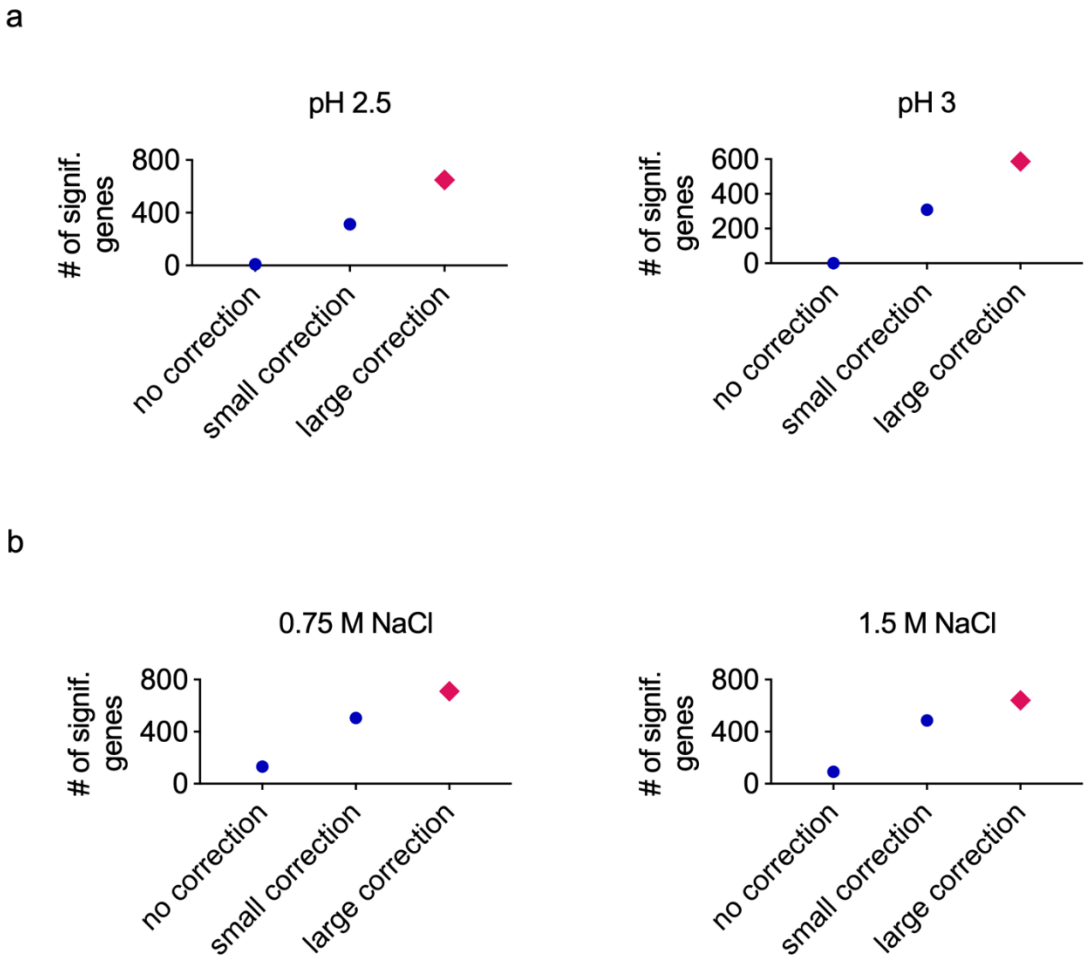

Figure S5.5. Number of significant genes at different levels of activity correction for low pH and high salt tolerance screens. Dark blue points represent the number of significant genes predicted by acCRISPR without CS correction and with a small CS correction (CS-threshold = 2.0), while pink diamonds indicate the number of predicted significant genes with a large CS correction (CS-threshold = 4.5, i.e., optimum CS-threshold) for (a) the two pH conditions, and (b) the two salt conditions.

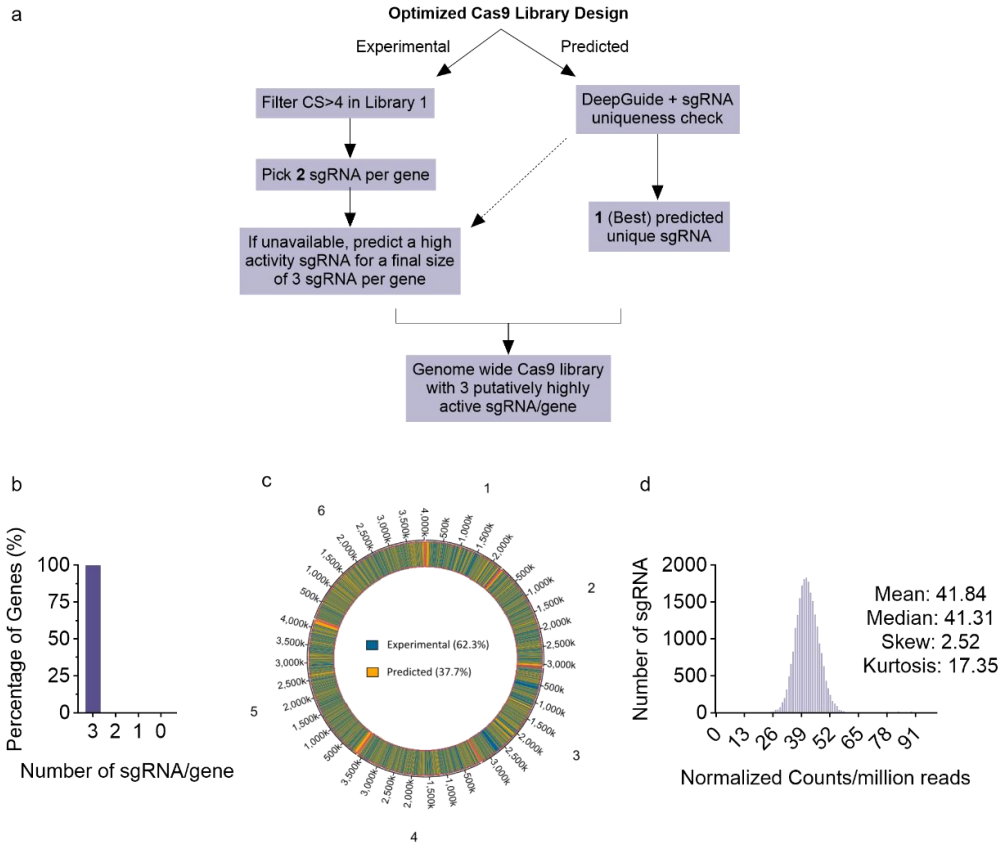

Figure S5.6. Design and characterization of the optimized Cas9 library in *Y. lipolytica*. (a) Flowchart of optimized library design. Every gene in the optimized library was designed to contain two experimentally validated high activity guides from the first Cas9 screen (CS>4.0) whenever possible, as well as one best predicted sgRNA by DeepGuide. All guides in the final library were verified to target a unique locus, as well as have unique seed sequences (11 nt closest to the PAM). (b) Per gene coverage of sgRNA. 99.84% or 7906 out of 7919 mRNA coding genes in Y. lipolytica were designed to have 3 sgRNA in the optimized library. Only 2 genes had no sgRNA designed. (c) Fraction of experimental and predicted sgRNA within the library. While the ideal design would have constituted 2/3rd and 1/3rd fractions of experimental and predicted sgRNA, the additional CS>4.0 filter applied to the experimental sgRNA limited the number of experimental validated sgRNA capable of being designed. These were supplemented by another predicted sgRNA with high activity such that the total number of guides per gene was 3 when possible. Thus, final library contained 62.3% and 37.7% experimental and predicted sgRNA. (d) A library consisting of 23,900 sgRNAs was synthesized by Agilent, cloned in-house and characterized by next generation sequencing. The library exhibited a tight normal distribution with nearly equal mean and median, and high kurtosis, signifying an even T-distribution.

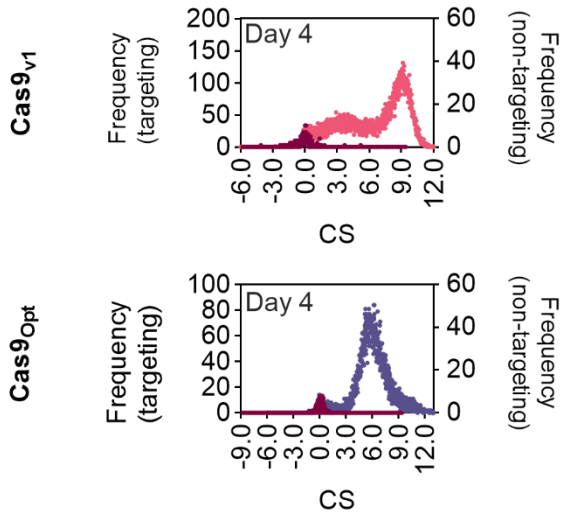

Figure S5.7. CS distributions of optimized and unoptimized CRISPR-Cas9 libraries with the nontargeting population mean normalized to 0. CS distributions were calculated on day 4. Purple and Pink distributions plotted on the left y-axis show CS values of optimized and the older unoptimized Cas9 libraries respectively, while the dark red data plotted with the right y-axis depicts the non-cutting control population, constituting ~1.5% of the respective library. The higher the value of CS, the better the cutting activity of the sgRNA. The mean of the nontargeting population at day 4 is normalized to 0 for both versions of the Cas9 library to visualize the spread of the activity profiles as compared with noncutters in both libraries.

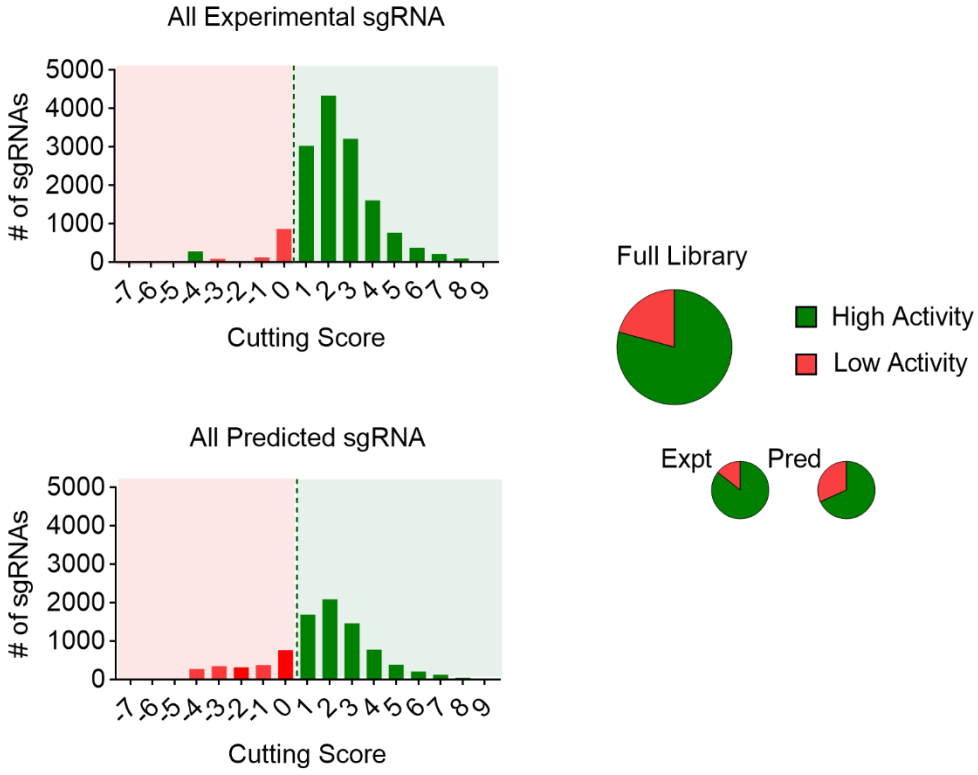

Figure S5.8. Characterization of sgRNA activity in the optimized Cas9 library. An sgRNA activity threshold of 1.0 was applied to the data denoted by the dashed line. Green and red rectangular regions on the plot represent CS values indicative of high and low guide activity. Green colored bars indicate that activity profiles of the guide matched with the expected design criteria. All predicted sgRNA were designed to be highly active, thus red bars signify guides that did not conform with this trend and instead fell into the low activity region. Most experimental guides were designed to be highly active, with the exception of 360 non targeting sgRNA designed to be inactive. These nontargeting sgRNA are represented by the small green bar in the low activity region of the frequency distribution of all experimental sgRNA. Pie charts indicative of the fraction of sgRNA that are highly active and poorly active for all experimentally validated, all predicted, and the total library is depicted on the right.

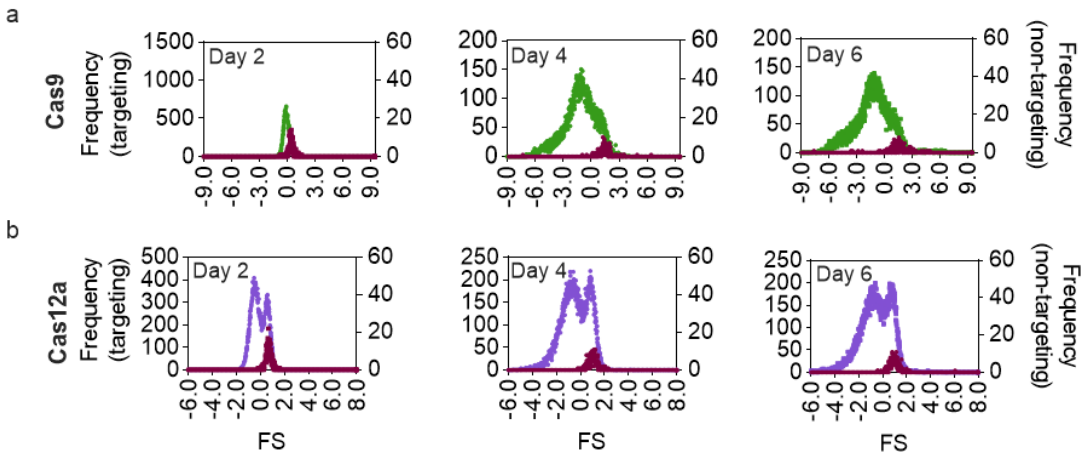

Figure S5.9. CRISPR-Cas9 and -Cas12a FS distributions on days 2, 4 and 6. Green and purple distributions plotted on the left y-axis show FS of all targeting sgRNA in the library, while the dark red distributions plotted on the right y-axis represents the non-targeting populations. (a) Histogram of sgRNA FS values in the Cas9 dataset. (b) Histogram of sgRNA FS values in Cas12a dataset.

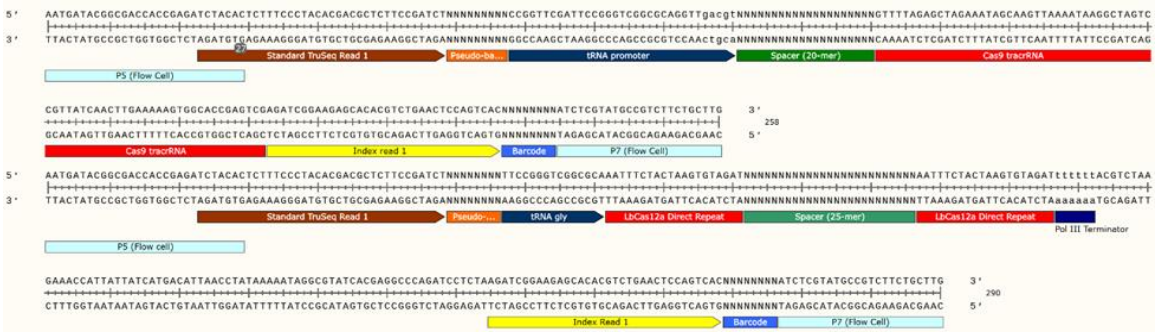

Figure S5.10. Schematic and sequence information of Cas9 (top) and Cas12a (bottom) amplicons for NGS. Amplicons contain: (i) P5 and P7 sequences (light blue) that are necessary for binding with the flow cell in Illumina sequencers, (ii) TruSeq adapter (brown) for binding of the sequencing primer, (iii) a portion of $tRNA^{gly}$ (black) expressing the sgRNA, (iv) Cas9 or Cas12 spacer (green) (v) Cas12a associated direct repeats or a portion of the Cas9 tracrRNA sequence (red), (vi) Universal 8 bp Illumina barcodes (blue), (vii) Index read 1 sequence for the binding of primers to sequence the Illumina barcodes, and (viii) 4-9 nt pseudo-barcodes (orange) at the 5' end between the TruSeq and $tRNA^{gly}$ which help demultiplex replicates that contain the same Illumina barcode.

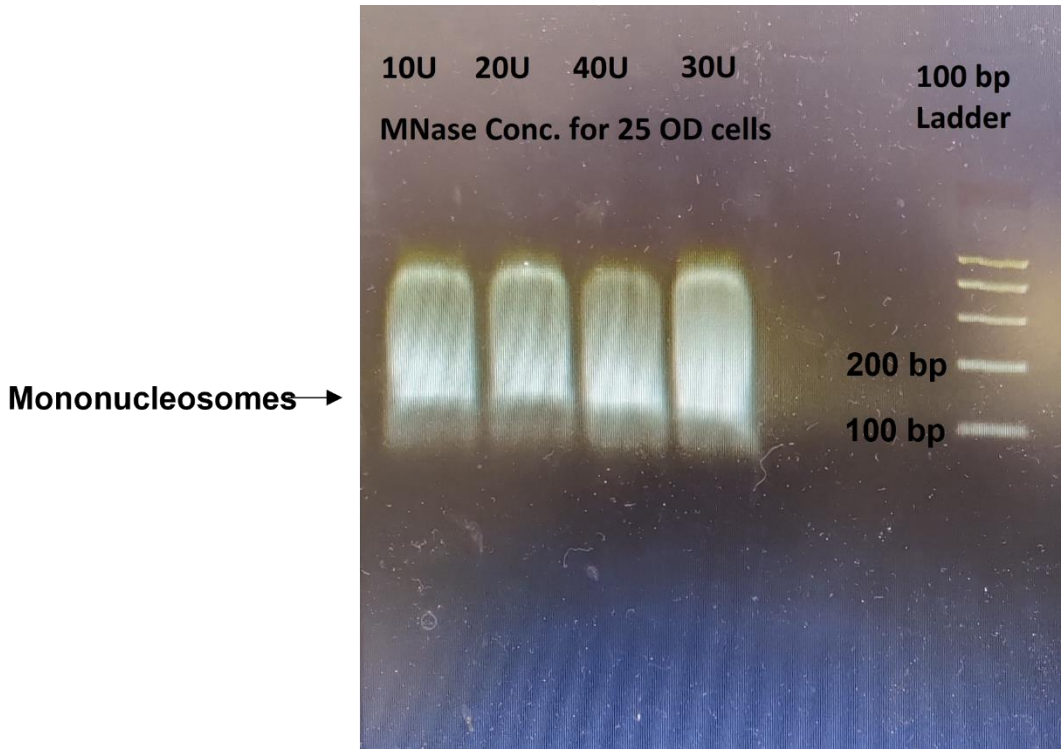

Figure S5.11. MNase titration for isolating mononucleosomal DNA. Micrococcal nuclease (MNase) concentrations of 10U, 20U, 30U and 40U per 25 OD of cells initially taken from exponential phase cultures, were tested. 100 ng of DNA was run for each well in the agarose gel pictures above and 40U of MNase gave the brightest bands of mononucleosomal DNA (~147 bp) with the least bands of trinucleosomal DNA (~450 bp).

5.11.2 Supplementary Tables

Table S5.1. CS threshold data for Cas9 and Cas12a screens. The CS threshold values used to generate 'CS-corrected' libraries and the optimum cutoff value for Cas9 and Cas12a datasets.

Cas9 Screen	Value			
Cutting efficiency score	Lowest cutoff	Highest cutoff	Step size	Optimum cutoff
Experimental CS	0.5	6.0	0.5	4.5
DeepGuide CS	0.5	6.0	0.5	4.0
Designer v1	0.108	0.892	0.098	0.402
Designer v2	20.209	78.441	7.279	49.325
CRISPRspec	1.215	39.175	4.745	15.45
CRISPRscan	0.491	0.739	0.031	0.553
SSC	0.301	0.789	0.061	0.484
uCRISPR	10.045	90.005	9.995	70.015

Cas12a Screen	Value			
Cutting efficiency score	Lowest cutoff	Highest cutoff	Step size	Optimum cutoff
Experimental CS	0.5	3.0	0.5	1.5
DeepGuide CS	0.5	2.5	0.5	1.0
DeepCpf1	10	90	10	40

Table S5.2. Yeast strains used in this study.

Yeast strain genotype	Phenotype
PO1f (MatA, *leu2-270*, *ura3-302*, *xpr2-322*, *axp-2*)	Wild type strain
PO1f *Δku70*	PO1f with disrupted KU70, which facilitates the non-homologous end joining DNA repair pathway
PO1f UAS1B8-TEF(136)-Cas9 -CycT::A08	PO1f expressing *Y. lipolytica* codon optimized Cas9 gene at the A08 locus
PO1f UAS1B8-TEF(136)-LbCas12a -CycT::A08	PO1f expressing *Y. lipolytica* codon optimized LbCas12a gene at the A08 locus
PO1f *Δku70* UAS1B8-TEF(136)-Cas9 -CycT::A08	*KU70* disrupted in Cas9 integrated PO1f strain
PO1f *Δku70* UAS1B8-TEF(136)-LbCas12a -CycT::A08	*KU70* disrupted in LbCas12a integrated PO1f strain

Table S5.3. Plasmids used for genome wide CRISPR screens.

Plasmid name	Reference	Function
pCpf1_yl	6	Plasmid for CRISPR-LbCas12a based gene editing in *Y. lipolytica*
pCRISPRyl (Addgene #70007)	7	Plasmid for CRISPR-Cas9 based gene editing in *Y. lipolytica*
pLbCas12ayl	This study and 4	Plasmid for CRISPR-LbCas12a based gene editing in *Y. lipolytic*a. sgRNA is flanked on either end by the direct repeat, to allow sgRNAs to end in T residues without being construed as part of the PolyT terminator
pHR_A08_hrGFP (Addgene #84615)	8	Plasmid containing homology arms for integration of hrGFP into the A08 locus
pHR_A08_LbCas12a	This study and 4	Plasmid containing homology arms for integration of LbCas12a into the A08 locus
pHR_A08_Cas9	9	Plasmid containing homology arms for integration of Cas9 into the A08 locus
pLbCas12ayl-GW	This study and 4	Vector containing sgRNA expression cassette for cloning Cas12a sgRNA library. (Does not contain Cas12a expression cassette)
pCas9yl-GW	9	Vector containing sgRNA expression cassette for cloning Cas9 sgRNA library. (Does not contain Cas9 expression cassette)
pCRISPRyl_KU70	This study and 10	CRISPR plasmid for the disruption of KU70

Table S5.4. Sequences of primers used in this study.

Primer name	Primer Sequence
ExtraDR-F	CGGCGCAAATTTCTACTAAGTGTAGACTAGTAATTTCTACTAAGTGTAGATTTTTTTACGTCTAAGAAACCATTATT
ExtraDR-R	AATAATGGTTTCTTAGACGTAAAAAAATCTACACTTAGTAGAAATTACTAGTCTACACTTAGTAGAAATTTGCGCCG
Cpf1-Int-F	TGCCTGGAGCCGAGTACGGCATTGATTACTAGTCCGGGTTCGAAGGTACCAAG
Cpf1-Int-R	TTAGGCTGGGTCTCGAGAGCAAAGAAGCCTAGGGCAAATTAAAGCCTTCGAGCG
BRIDGE-F	CTAAATTTGATGAAAGGGGGATCCCCCGGGTGGCGTAATCATGGTCATAGCTGTTTCCTG
BRIDGE-R	CAGGAAACAGCTATGACCATGATTACGCCACCCGGGGGATCCCCCTTTCATCAAATTTAG
A08-Seq-F	AGCCGAGTACGGCATTGAT
A08-Seq-R	TCAATGTAGCCTCCTCCAACC
Tef_Seq-F	GTTGGGACTTTAGCCAAG
Lb1-R	CTTCTGCTTGGTCTTCTGGTTG
Lb2-F	AACCTGTACAACCAGAAGACCAAG
Lb3-F	AAGGAGACCAACCGAGACGAG
Lb4-F	AACCTGCACACCATGTACTTCAAG
Lb5-F	CCAGATCACCAACAAGTTCGAGTC
M13-F	GTAAAACGACGGCCAGT
InversePCR-F	TTTTTTTACGTCTAAGAAACCATTATTATCATGACATTAACCT
InversePCR-R	TGCGCCGACCCGGAATCGAACCGGGGGCCC
OLS-F	GTTTAGTGGTAAAATCCATCGTTGCCATCG
OLS-R	GATACGCCTATTTTTATAGGTTAATGTCATG
qPCR-GW-F	TTATGAACTGAAAGTTGATGGC
qPCR-GW-R	TCACACAGGAAACAGCTATG
Cr_1250	TATAAGAATCATTCAAAGGCGCGCATGGATAAGAAATACTCCATTGGCCTG
Cr_1254	ATAACTAATTACATGAGGCTAGCTTACAGCATGTCCAGATCGAAATCG

Table S5.5. Transformation efficiencies measured as x10^6 transformants, for all replicates in the control and treatment strains.

Cas9 Screen	Replicate Transformation Efficiency (x10^6 transformants)		
Strain	R1	R2	R3
PO1f	12.35	11.39	15.80
PO1f Cas12a	11.42	8.29	10.64
PO1f Cas12a Δku70	6.79	7.33	7.08

Cas12a Screen	Replicate Transformation Efficiency (x10^6 transformants)		
Strain	R1	R2	R3
PO1f Δku70	6.89	6.21	5.43
PO1f Cas12a Δku70	5.06	4.29	4.41
PO1f	11.93	8.28	4.23
PO1f Cas12a	6.32	5.47	6.11

Table S5.6. Primers used for NGS fragment amplification (Cas12a)

Primer name	Primer Sequence	Illumina Barcode (Reverse primer) / Pseudo-Barcode (Forward primer) for demultiplexing
ILU1-F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTTCCGGGTCGGCGCAAATTTC	^TTCCGG
ILU2-F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGATCGGGTCGGCGCAAATTTCT	^AGATCG
ILU3-F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGCTATTCGGGTCGGCGCAAATTTCT	^GCTATT
ILU4-F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGGACTACGGGTCGGCGCAAATTTCT	^CAGGAC
ILU1-R	CAAGCAGAAGACGGCATACGAGATTCGCCTTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	CAAGGCGA
ILU2-R	CAAGCAGAAGACGGCATACGAGATGACGAGAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	CTCTCGTC
ILU3-R	CAAGCAGAAGACGGCATACGAGATAGACTTGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	CCAAGTCT
ILU4-R	CAAGCAGAAGACGGCATACGAGATCTGTATTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	TAATACAG
ILU5-R	CAAGCAGAAGACGGCATACGAGATCCTGAACCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	GGTTCAGG
ILU6-R	CAAGCAGAAGACGGCATACGAGATATCAGGTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	AACCTGAT
ILU7-R	CAAGCAGAAGACGGCATACGAGATTAGGTGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	GTCACCTA
ILU8-R	CAAGCAGAAGACGGCATACGAGATCGAACAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	ACTGTTCG
ILU9-R	CAAGCAGAAGACGGCATACGAGATGTTCGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	GATCGAAC
ILU10-R	CAAGCAGAAGACGGCATACGAGATACCTAGCTGTGACTGGAGTTCAGACGTGTGCCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	AGCTAGGT
ILU11-R	CAAGCAGAAGACGGCATACGAGATAGAGATGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	TCATCTCT
ILU12-R	CAAGCAGAAGACGGCATACGAGATCTGGACTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	AAGTCCAG

Table S5.7. Primers used for NGS fragment amplification (Cas9)

Primer name	Primer Sequence	Illumina Barcode (Reverse primer) / Pseudo-Barcode (Forward primer) for demultiplexing
Cr_1665	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG ACGCTCTTCCGATCTAGTCCGGTTCGATTCCGGGTC	^AGTCCG
Cr_1666	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG ACGCTCTTCCGATCTGTAGTCCGGTTCGATTCCGGGTC	^GTAGTC
Cr_1667	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG ACGCTCTTCCGATCTCAGTAGTCCGGTTCGATTCCGGGTC	^CAGTAG
Cr_1668	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACG ACGCTCTTCCGATCTTCCAGTAGTCCGGTTCGATTCCGGGTC	^TCCAGT
Cr_1669	CAAGCAGAAGACGGCATACGAGATTCGCCTTGGTGACTGGAG TTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTC AAG	CAAGGCGA
Cr_1670	CAAGCAGAAGACGGCATACGAGATATAGCGTCGTGACTGGAG TTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTC AAG	GACGCTAT
Cr_1671	CAAGCAGAAGACGGCATACGAGATGAAGAAGTGTGACTGGAG TTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTC AAG	ACTTCTTC
Cr_1672	CAAGCAGAAGACGGCATACGAGATATTCTAGGGTGACTGGAG TTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTC AAG	CCTAGAAT
Cr_1673	CAAGCAGAAGACGGCATACGAGATCGTTACCAGTGACTGGAG TTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTC AAG	TGGTAACG
Cr_1709	CAAGCAGAAGACGGCATACGAGATGTCTGATGGTGACTGGAG TTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTC AAG	CATCAGAC
Cr_1710	CAAGCAGAAGACGGCATACGAGATTTACGCACGTGACTGGAG TTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTC AAG	GTGCGTAA
Cr_1711	CAAGCAGAAGACGGCATACGAGATTTGAATAGGTGACTGGAG TTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTC AAG	CTATTCAA

Table S5.8. Parameters for bioinformatics tools on Galaxy [11] used in the analysis of NGS reads (Cas12a)

Tool	Version	Parameters*
FastQC	v0.11.8	Default settings
Cutadapt	Galaxy Version 1.16.6 [12]	The 3 biological replicates of a given sample at a given time-point in the Cas12a screen always had the same reverse primer containing the Illumina barcode, and forward primers ILU1-F, ILU3-F and ILU4-F; or ILU2-F, ILU3-F and ILU4-F each containing different pseudo-barcodes. Thus Cutadapt was used to demultiplex biological replicates from each other. • 5' (Front) anchored 6 bp pseudo-barcodes to be demultiplexed (-g): ^NNNNNN (refer to previous table for pseudo-barcode-forward primer association). • Maximum error rate (--error-rate): 0.2 • Match times (--times): 1 • Minimum overlap length (--overlap): 4 • Multiple output: Yes (Each demultiplexed readset is written to a separate file)
Trimmomatic	v0.38	• HEADCROP: 29 (if amplified by ILU1-F); or 30 (if amplified by ILU2-F); or 32 (if amplified by ILU3-F); or 34 (if amplified by ILU4-F) • CROP: 25
Bowtie2**	v2.4.2	• Number of allowed mismatches in seed alignment (-N): 1 • Length of the seed substring (-L): 21 • Function governing interval between seed substrings in multiseed alignment (-i): S,1,0.50 • Function governing maximum number of ambiguous characters (--n-ceil): L,0,0.15 • Alignment mode: end-to-end • Number of attempts of consecutive seed extension events (-D): 20 • Number of times re-seeding occurs for repetitive reads: 3 • Save mapping statistics: Yes

* All parameters other than those mentioned here are kept at default values.
** Bowtie2 usage needs a genome fasta file for alignment. Nontargeting sgRNA and any other sgRNA that Bowtie2 could not find within the original CLIB89 genome file were appended as an extra chromosome so that Bowtie could align all sgRNA for the purposes of generating counts.

Table S5.9. Parameters for bioinformatics tools on Galaxy [11] used in the analysis of NGS reads (Cas9)

Tool	Version	Parameters*
FastQC	v0.11.8	Default settings
Cutadapt	Galaxy Version 1.16.6 [12]	Cutadapt was used to demultiplex samples containing the same Illumina barcode, but different pseudobarcodes at the 5' end of the read. Samples were amplified with reverse primers Cr1669-1673;Cr1709-1711 and forward primers Cr1665-1668 each containing a different pseudo barcode as mentioned in Table • 5' (Front) anchored 6 bp pseudo-barcodes to be demultiplexed (-g): ^NNNNNN (refer to previous table for pseudo-barcode-forward primer association). • Maximum error rate (--error-rate): 0.2 • Match times (--times): 1 • Minimum overlap length (--overlap): 4 • Multiple output: Yes (Each demultiplexed readset is written to a separate file)
Trimmomatic	v0.38	• HEADCROP: 30 (if amplified by Cr1665); or 32 (if amplified by Cr1666); or 34 (if amplified by Cr1667); or 36 (if amplified by Cr1668) • CROP: 20
Bowtie2**	v2.4.2	• Number of allowed mismatches in seed alignment (-N): 1 • Length of the seed substring (-L): 19 • Function governing interval between seed substrings in multiseed alignment (-i): S,1,0.50 • Function governing maximum number of ambiguous characters (--n-ceil): L,0,0.15 • Alignment mode: end-to-end • Number of attempts of consecutive seed extension events (-D): 20 • Number of times re-seeding occurs for repetitive reads: 3 • Save mapping statistics: Yes

* All parameters other than those mentioned here are kept at default values.
** Bowtie2 usage needs a genome fasta file for alignment. Nontargeting sgRNA and any other sgRNA that Bowtie2 could not find within the original CLIB89 genome file were appended as an extra chromosome so that Bowtie could align all sgRNA for the purposes of generating counts.

5.11.3 References

1. Allen, F. *et al.* JACKS: joint analysis of CRISPR/Cas9 knockout screens. *Genome Res.* **29**, 464–471 (2019).

2. Li, W. *et al.* Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biol.* **16**, 281 (2015).

3. Daley, T. P. *et al.* CRISPhieRmix: a hierarchical mixture model for CRISPR pooled screens. *Genome Biol.* **19**, 159 (2018).

4. Baisya, D., Ramesh, A., Schwartz, C., Lonardi, S. & Wheeldon, I. Genome-wide functional screens enable the prediction of high activity CRISPR-Cas9 and -Cas12a guides in Yarrowia lipolytica. *Nat. Commun.* **13**, 922 (2022).

5. Luo, J., Chen, W., Xue, L. & Tang, B. Prediction of activity and specificity of CRISPR-Cpf1 using convolutional deep learning neural networks. *BMC Bioinformatics* **20**, 332 (2019).

6. Ramesh, A., Ong, T., Garcia, J. A., Adams, J. & Wheeldon, I. Guide RNA Engineering Enables Dual Purpose CRISPR-Cpf1 for Simultaneous Gene Editing and Gene Regulation in. *ACS Synth. Biol.* **9**, 967–971 (2020).

7. Schwartz, C. M., Hussain, M. S., Blenner, M. & Wheeldon, I. Synthetic RNA Polymerase III Promoters Facilitate High-Efficiency CRISPR-Cas9-Mediated Genome Editing in Yarrowia lipolytica. *ACS Synth. Biol.* **5**, 356–359 (2016).

8. Schwartz, C., Shabbir-Hussain, M., Frogue, K., Blenner, M. & Wheeldon, I. Standardized Markerless Gene Integration for Pathway Engineering in Yarrowia lipolytica. *ACS Synth. Biol.* **6**, 402–409 (2017).

9. Schwartz, C. *et al.* Validating genome-wide CRISPR-Cas9 function improves screening in the oleaginous yeast Yarrowia lipolytica. *Metab. Eng.* **55**, 102–110 (2019).

10. Schwartz, C., Frogue, K., Ramesh, A., Misa, J. & Wheeldon, I. CRISPRi repression of nonhomologous end-joining for enhanced genome engineering via homologous recombination in Yarrowia lipolytica. *Biotechnol. Bioeng.* **114**, 2896–2906 (2017).

11. Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.* (2022) doi:10.1093/nar/gkac247.

12. Jalili, V. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.* **48**, W395–W402 (2020).

Chapter 6: Summary and prospective future directions

Y. lipolytica's rise to industrial favor has been facilitated by a combination of the attractive phenotypes it presents, and the development of synthetic biology tools that helps leverage these traits. While there now exist capabilities for basic genome editing and transcriptional regulation, there is a need for new tools and workflows for multiplexed genome editing, and forward genetic screening. The work presented in this dissertation addresses those needs and focuses on expanding the available toolset for genome engineering and novel gene discovery, to accelerate design-build-test-learn-cycles for strain engineering. We started by adapting the now widely used CRISPR-Cas12a system from *Lachnospiraceae bacterium* (LbCas12a) as an orthogonal supplement to the existing Cas9 toolset. We also demonstrated LbCas12a's ability to process multiple tiled spacers into mature sgRNA by simultaneously knocking out three genes on different chromosomes, with high efficiency.

The spacer length dependent cutting activity of Cas12a was thoroughly investigated and characterized, and the loss of cutting (while still binding to genomic target) at shorter spacer lengths was leveraged to expand CRISPRi and CRISPRa modalities to Cas12a. Standard techniques for CRISPR based transcriptional control involve the fusion of transcriptional regulator proteins to a nuclease inactive mutant of the Cas endonuclease (dCas), to then be targeted upstream of a desired gene's transcription start site. An alternative method to catalytic deactivation, where Cas nuclease activity could instead be

modulated based on spacer length, was demonstrated here. Combining Cas12a's ability to multiplex, with these expanded transcriptional regulation modalities, it was also shown that simultaneous gene disruption and silencing at two different loci was also possible.

While the developed a CRISPR-Cas12a system in *Y. lipolytica* has enabled rapid strain development through simultaneous gene disruptions, there exist further exciting avenues to be explored. Gene integration into the genome is a preferred strategy for overexpression of any pathway genes as plasmid-based expression is more unstable with cell-to-cell variability in culture. Typical strategies for gene integration make use of the native homologous recombination pathway for DSB repair, however *Y. lipolytica* preferentially makes use of non-homologous end joining to repair DNA (as do most non-conventional yeasts), which adds a layer of complexity. As well, integrations are typically facilitated by auxotrophy or antibiotic resistance markers that need to be removed at each step, further slowing the process. To address this issue, a previous work characterized five standardized sites in the genome amenable for efficient gene integration and developed a markerless integration strategy using the CRISPR-Cas9 system. However, there is still variability in integration rates with this method, especially with longer insert lengths. Since pathway engineering typically requires co-expression of multiple genes, more than one gene would need to be integrated into a specific site, increasing insert lengths further. As such there exist possible avenues for improvement in this area with the developed CRISPR-Cas12a system.

Class 2 Type V endonucleases like Cas12a create staggered DSB with sticky ends unlike Cas9 which creates blunt end breaks. It has previously been shown in plants, certain algae and in human cell lines, that having DNA overhangs increase rates of homologous recombination. Thus, adapting the markerless integration strategy to the Cas12a system could improve gene integration efficiency in *Y. lipolytica*. Further, it was previously demonstrated that repressing the NHEJ (specifically the KU70 and KU80 genes) machinery using CRISPRi, resulted in improved rates of homology directed gene integration. Combining Cas12a's tendency to create DSB with overhangs, with its demonstrated ability to perform multiplexed gene disruption and silencing, KU70 could be repressed at the same time as performing a targeted integration, thereby improving the chances of success. Other target genes for silencing could include those involved in cell cycle regulation. It is known that increasing the time spent in G2 or S phase of the cell cycle increases homologous recombination events. Thus, repressing genes involved in progressing the cell cycle from the interphase could theoretically improve HDR.

Orthogonality of various Cas proteins due to vastly different PAM requirements open the possibility of multiplexing CRISPRa, CRISPRi, and CRISPR-KO modalities for combinatorial metabolic engineering. This has already been demonstrated in *S. cerevisiae* with the help of 3 orthogonal endonucleases SpCas9 (CRISPRi), LbCas12a (CRISPRa), and SaCas9 (CRISPR-KO). Simultaneous modulation by overexpression of rate limiting enzymes with CRISPRa, and repression or deletion of genes that divert flux into competing pathways (repressing essential genes; disrupting non-essential genes) will help speed up testing and screening cycles for strain building. *Y. lipolytica* displays many attractive

phenotypes like halotolerance, pH tolerance and natively silenced genes for the consumption of pentose sugars. These traits are a result of complex genetic interactions and will benefit from combinatorial engineering strategies for their enhancement. CRISPR technologies like those described above would enable the investigation of gain-of-function and loss-of-function combinations that synergistically enhance these traits without the need for cloning new strains with different promoters.

Pooled CRISPR screens offer the ability to unbiasedly interrogate the role of gene function in relation to a specific phenotype. Use of such screens in non-model organisms have been limited, in part due to the inability to predict and design highly active sgRNA that are essential for accurate screening. In Chapter 4, this issue was addressed by the design of a deep learning model called DeepGuide that could predict highly active sgRNA for Cas9 and Cas12a based CRISPR systems in *Yarrowia*. A Cas12a sgRNA library that covered nearly every protein coding gene in the genome with eight-fold redundancy was created to supplement the Cas9 library. Negative selection screens in a strain deficient in DNA repair generated guide activity profiles which was further used to train DeepGuide with examples of highly active and poorly active guides. While Cas nucleases themselves have certain nucleotide preferences for gRNA (for e.g., disfavoring thymines and instead favoring guanine or cytosines in the PAM proximal region for Cas9; avoiding spacer-direct repeat complementarity in Cas12a), guide activity is also dictated by epigenetic features such as chromatin accessibility or DNA methylation in CpG islands. DNA methylation to 5-methylcytosine is said to occur at very low frequencies in *Yarrowia* (<0.5% compared to over 4% in human genomes) and is thus unlikely to influence guide activities to a large

extent. Meanwhile, nucleosome positioning can sterically hinder access of Cas9/Cas12a to the target locus, thereby affecting nuclease activity. Thus, *Yarrowia* nucleosome positioning data was included and found to improve gRNA activity predictions. Other parameters that improved the performance of the algorithm included the size of the library that the algorithm was trained on, as well as genomic context surrounding an sgRNA. Larger library sizes for training and including the few nucleotides present on either end of the sgRNA improved activity predictions.

A current open challenge in machine learning based guide design is the lack of cross species predictive capability. Chapter 4 also discusses how tools designed for guide prediction in mammalian cells do not perform well in *Yarrowia* and vice versa. Given the role of epigenetic features as a determinant of Cas nuclease activity, sgRNAs that perform well in one organism may not perform well in others. One possible approach to this problem is compiling guide activity data from the growing list of CRISPR screens (e.g., databases such as BioGRID and iCSDB) across all organisms to fully establish Cas specific guide requirements and then overlaying organism specific features such as chromatin accessibility, and DNA and histone methylation on a case-by-case basis. However, even Cas specific design rules are still being uncovered and further research will be required to achieve true cross species predictivity.

Yet another challenge in CRISPR guide design lies in its application to transcriptional regulation through CRISPRi or CRISPRa. As discussed previously, the primary purpose of these sgRNA is not to induce nuclease activity of the Cas protein, but merely to ensure

tight binding to the target region so that the fused transcriptional regulators may perform their role. While guide activity for gene editing is easily characterized, detailed characterizations probing the link between guide sequence and transcriptional control have not yet been established. Furthermore, the narrow spatial window upstream of the start codon for effective transcriptional regulation, further complicates guide design.

Decreasing costs of DNA synthesis and sequencing may facilitate the investigation of this problem. For example, a fluorescent protein like GFP or dsRed may first be episomally expressed from a small, synthesized library of native promoters (low expression promoters for testing gene activation, and high expression promoters for testing gene repression). Subsequently, a library of gRNA for CRISPRa/i, targeting each promoter may be combinatorially cloned into plasmids containing the cognate target promoter. Evaluation of cutting activity for each sgRNA may simply be measured by plasmid loss assays in a strain containing Cas and deficient in DNA repair. Meanwhile, evaluation of transcriptional control may be measured by fluorescence assays in a strain expressing dCas. Such an experiment would provide a rich dataset with which to correlate guide cutting activity to its role in transcriptional control.

The last chapter of this dissertation focuses on the application of the constructed genome wide Cas9 and Cas12a libraries to solve a biological problem in *Y. lipolytica*, namely the definition of a consensus set of essential genes. With the dearth of pooled CRISPR screens in non-conventional hosts, comes a lack of specific analysis workflows available to determine accurate screening hits. Chapter 5 emphasizes this point by showing

that existing essential gene prediction tools developed from mammalian cell screening datasets are unable to accurately capture *Yarrowia* screening results. We address this issue with the development of acCRISPR, an end-to-end analysis workflow capable of taking gRNA read counts generated from NGS data and providing a list of genes essential to the screening condition. We then used acCRISPR analysis of Cas9 and Cas12a, along with results from a previous transposon-based screen, to arrive at consensus a set of essential genes for *Yarrowia's* growth on glucose. Not only were the number of essential genes predicted in the range of what is expected in yeasts (based on published analyses in *S. cerevisiae*, *S. pombe* and *R. toruloides*), but GO analysis further showed enrichment of essential biological processes such as transcription, translation, cell cycle regulation and ribosome biogenesis, lending further confidence to the results.

acCRISPR's accuracy came from its use of an additional experimental dataset to determine gRNA activities and then using these to provide activity correction to gene fitness scores. This step is critical in ensuring that poorly active guide RNAs targeting essential genes do not obscure their essentiality. While this offers an undeniable advantage in improving hit calling accuracy, we also acknowledge that such experiments may not be easy to conduct in all organisms. Thus, we also allowed acCRISPR to make use of predicted gRNA activities from available tools, in place of experimental scores. Such an analysis for the *Yarrowia* dataset, showed mediocre performance when predicted sgRNA activities from a few well-known tools trained on mammalian cell data were used. However, essential gene prediction on DeepGuide predicted guide activities surpassed those of other activity prediction tools, even if it did not surpass the experimental results.

It is important to note that the prediction tools used all showed superior performance in the species they were trained in and that they could not be expected to accurately perform in a species that they were not exposed to. As organism-specific guide activity predictions improve, we may expect them to slowly take the place of experimental scores.

We also utilized acCRISPR in the analysis of loss- and gain-of-fitness screens to enhance *Yarrowia's* native tolerance to salt and acidic pH. As discussed in chapter 1, halo and osmotolerance can enhance bioprocess economics by allowing for the use of cheaper water sources like seawater, while tolerance to highly acidic pH may help reduce sterilization costs by enabling non-aseptic culturing. In *Y. lipolytica*, lipid biosynthesis competes with the production of citrate for the precursor acetyl-CoA, and other studies suggest that low pH in the culture media can inhibit secretion of citrate, and improve its conversion back to acetyl-CoA, allowing for better lipid yields. Our CRISPR screens identified known and novel gene hits that were either critical for tolerance to salt or pH, or which when disrupted conferred improved tolerance. In the case of halotolerance, a gene that had homology to ROT2 (glucosidase subunit) in *S. cerevisiae* was classified as a gain-of-fitness hit. Knockout of ROT2 has previously shown improved chitin content of the cell was in *S. cerevisiae*, and it is possible that this confers increased durability under salt stress. Similarly, the knockout of the top gain-of-fitness hit for tolerance to acidic pH has been implicated in conferring acid tolerance in *S. cerevisiae* as well. Thus, this exercise demonstrates the utility of CRISPR screens in the unbiased interrogation of genetic determinants to interesting phenotypes.

Finally, we use our guide activity scores obtained from the previous CRISPR screens, as well as DeepGuide's guide activity predictions to design an optimized minimal Cas9 library. The motivation behind such an endeavor was once again to limit the effect of poorly active sgRNA on determining gene fitness effects. This smaller size of this library also had the added advantage of limiting the transformation efficiency burden inherent in pooled screens. Pooled screens typically require a minimum transformation efficiency of a 100-fold to even 500-fold or more transformants in comparison to the library size. Typical guide design strategies involve designing many guides per gene to offset the effect of poorly active guides, unnecessarily bloating the library size. Our optimized Cas9 library targeted every mRNA coding gene in the Yarrowia genome, with putatively three high activity guides per gene. Experimental validation of the library suggested that over 80% of the library was indeed highly active. Further, in the case of the first Cas9 library, acCRISPR eliminated over half sgRNA (>23,000 sgRNA) as poorly active for the purposes of accurate essential gene prediction. Meanwhile, similar analysis with the optimized library indicated with the elimination of small fraction of sgRNA (~4,000 sgRNA) the optimized library is capable of accurate essential gene predictions.

The development of an optimized library opens possibilities of answering many other interesting biological questions in *Y. lipolytica*. While lipid metabolism has been relatively well characterized in *Yarrowia* through rational metabolic engineering, there likely exist non-obvious hits for improving lipid production. SNF2 is one such example of a gene unrelated to storage lipid biosynthesis, that was previously identified to improve lipid accumulation in *S. cerevisiae* through transposon screening. This disruption of this gene

was also found to improve TAG biosynthesis in *Yarrowia* due to its role as a regulator of the gene ACC1. ACC1 catalyzes the first committed step fatty acid biosynthesis and is tightly controlled in *Yarrowia* through the action of kinases (such as SNF2) that phosphorylate the Acc1p to abolish its catalytic activity.

It is thus possible to identify other novel genes that play a role in lipid accumulation through a genome wide screen. Conducting these screens in the presence of a selection pressure like the addition of the chemical cerulenin is also likely to improve screening results. Cerulenin is an antifungal agent whose activity interferes with the formation of fatty acid synthesis, which limits the formation of essential membrane lipids. Mutants of *Yarrowia* that accumulate high levels of lipids despite cerulenin presence have likely rewired their metabolic pathways to combat cerulenin inhibition of fatty acid synthesis. Such screens may thus lead to the identification of novel knockout targets for improving lipid biosynthesis and are currently the subject of further study.

Another ongoing study involving the use of the optimized Cas9 library looks to identify novel gene hits that can improve *Yarrowia's* tolerance and utilization of acetate as a carbon source. When glucose (6C) is utilized as a carbon source, the glycolysis pathway breaks it down into 2 molecules of pyruvate (3C), which is further broken down to 2 molecules of acetyl-CoA (2C; 2 carbon equivalents lost as CO_2 in the entire cycle) by the pyruvate decarboxylase mechanism or the pyruvate dehydrogenase complex. On the other hand, when acetate (2C) is utilized as a carbon source, it is directly converted to acetyl-CoA (2C) via an acetyl-CoA synthetase with no loss in carbon equivalents. Given that

acetyl-CoA is the most important precursor shunted towards the production of many value-added compounds in *Y. lipolytica*, growth on acetate is a favorable trait to engineer in this organism. A recent study published in Nature Catalysis showed the ability to convert CO_2 into high values C_{2+} compounds, and more specifically acetate with high specificity. Thus, microorganisms that can use acetate from such a feed to produce high value compounds would be highly desirable for efficient bioprocess economics.

As underscored several times in this dissertation, non-conventional organisms can show a broad range of interesting phenotypes, that can be leveraged for the economic and scalable production of high value biochemicals and bioproducts. Domestication of these organisms has been greatly improved by advances in DNA synthesis, sequencing, and CRISPR based genome editing technologies. Significant effort has been invested in the development of tools and techniques to engineer these organisms into suitable production hosts. The work presented here as a whole contributes towards that goal by building advanced synthetic biology tools and computational analysis workflows for improving engineering efforts in *Y. lipolytica*.