Title

Analysis of Functional Genetic Screens for Genome-Wide Metabolic Engineering of Microbial Bioproduction Hosts

Author

Trivedi, Varun

Publication Date

2024

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Analysis of Functional Genetic Screens for Genome-Wide Metabolic Engineering of Microbial Bioproduction Hosts

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

Doctor of Philosophy

in

Chemical and Environmental Engineering

by

Varun Trivedi

September 2024

Dissertation Committee:
Dr. Ian Wheeldon, Chairperson
Dr. Robert Jinkerson
Dr. Stefano Lonardi

The Dissertation of Varun Trivedi is approved:

____

____

____
Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

As I reflect on the five year-long journey that led to this dissertation, I would like to express my deepest thanks to many individuals for their support, guidance, and encouragement. To begin with, I would like to express my sincere gratitude to my PhD advisor, Dr. Ian Wheeldon, for being supportive to my needs, being open to my research ideas, and fostering independence by trusting me to manage my tasks. I am grateful to him for helping me sharpen my communication skills as a researcher by pushing me to document my work into weekly reports, make appealing graphics to convey my research findings, and more importantly, present my research work as an engaging story to the audience – a highly rewarding but underrated skill in the field of research.

I would like to thank my lab colleagues, former and present, for being great friends at work, and am grateful for all the insightful research-related discussions as well as disagreements that we have had over these years. I am especially indebted to Adithya for patiently answering my countless biology-related questions as I was navigating this relatively unfamiliar field, and for mentoring me and making me feel comfortable when I first joined the lab. I am thankful to Mengwan for always being helpful and a good friend outside of work.

My deepest gratitude to Adithya Ramesh and Sangcheon Lee, who performed experiments and co-first authored the work described in Chapter 2. I am sincerely thankful to Sarah Thorwall for generating the experimental-computational data and being my co-first author for the work described in Chapter 4. I am also grateful for the support of my co-authors Adithya Ramesh, Aida Tafrishi, Cory Schwartz, Mengwan Li, Amirsadra Mohseni, and Eva Ottum, regarding the work described in Chapters 1, 2, 3 and 4. I extend my gratitude to Dr. Robert Jinkerson and Dr. Stefano Lonardi for serving on my dissertation committee, and for their valuable insights on my research during my oral qualifying exam. I am also thankful to Dr. Stefano Lonardi for being an amazing collaborator and co-supervising the work described in Chapter 3.

The text of Chapter 1 of this dissertation is a reprint of the material as it appears in the Journal of Industrial Microbiology and Biotechnology (2023). The text of Chapter 2 is a reprint of the material as it appears in the journal Communications Biology (2023). The text of Chapter 4 is a reprint of the material as it appears in the journal Metabolic Engineering (2023). The research work described in each article was supervised by the corresponding author, Dr. Ian Wheeldon. I sincerely thank UC Riverside, National Science Foundation, the US Department of Energy, Air Force Office of Scientific Research, and Army Research Office for supporting and funding the work described in this dissertation.

I am incredibly grateful to my parents and grandparents for their constant love, care, motivation, and for believing in my ability to pursue a PhD. They have been instrumental in looking after my well-being and this journey would not have been possible without their unrelenting support. I also sincerely appreciate my aunt, Bhoomika, for always being there for me, appreciating me for who I am, and wishing the best for me. My heartfelt gratitude to my family friends, Jyoti aunty and Aaditya, for helping me settle and feel homely in a foreign country, and treating me like a family member by ensuring that I am always taken care of. Last but not least, I am profoundly thankful to all my friends, especially Devesh, Nisarg and Yash, for regularly keeping in touch, always having my back, and having fun conversations with me on various topics that would lighten me up and energize me to keep working.

ABSTRACT OF THE DISSERTATION

Analysis of Functional Genetic Screens for Genome-Wide Metabolic Engineering of Microbial Bioproduction Hosts

by

Varun Trivedi

Doctor of Philosophy, Graduate Program in Chemical and Environmental Engineering
University of California, Riverside, September 2024
Dr. Ian Wheeldon, Chairperson

Microorganisms found in the environment around us exhibit a multitude of desirable characteristics that make them suitable hosts for the industrial production of biochemicals and biofuels. A common strategy to engineer microbes is using rational design approaches that entail manipulation of genes involved in one or more native metabolic pathways to improve biochemical synthesis or other relevant phenotypes. The advent of synthetic biology tools such as CRISPR-Cas systems for gene editing, and next-generation sequencing technologies has, however, made it possible to elucidate pangenome-wide targets for strain engineering by performing experiments at a genomic scale, such as genome-wide pooled CRISPR knockout screens, and analyzing the resulting high-throughput data using bioinformatic methods. A crucial challenge in evaluating the outcomes of pooled CRISPR screens is accounting for the variability in sgRNA knockout efficiency, as low-activity guides can potentially mask screening hits to result in false negatives. Towards that end, we developed an analysis method, acCRISPR, that processes NGS data from pooled CRISPR knockout screens, and provides an activity correction to accurately call statistically significant genes for the phenotype under study. We applied acCRISPR to CRISPR-Cas9 and CRISPR-Cas12a screening datasets from the oleaginous yeast *Yarrowia lipolytica* to identify a high-confidence set of essential genes for growth on glucose, as well as genes important for providing tolerance to high salt stress conditions. We further used the experimental sgRNA activity profiles from these screens to determine *in silico* sgRNA activity prediction accuracy of deep learning-based models trained on balanced and imbalanced experimental datasets, and improve prediction power with imbalanced training datasets by augmenting them with synthetic sgRNA. In another study, we sought to identify genetic targets responsible for phenazine biosynthesis in the bacterium *Pseudomonas chlororaphis* by employing a population genomics approach. We sequenced 34 Pseudomonas isolates using short- and long-read sequencing technologies, characterized them for phenazine production, and performed a microbial genome-wide association study (mGWAS) on the genomic-phenotypic data to elucidate the most influential phenazine biosynthesis targets across the pangenome. Overall, this work demonstrates the utility of high-throughput experimental-computational frameworks for identifying microbial strain engineering targets at a genomic scale and establishing novel genotype-phenotype relationships.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table S4.5. List of plasmids used in this study.

Chapter 0: Thesis organization

Functional genomic screening using various synthetic biology tools results in the generation of vast amounts of data, necessitating computational analysis and modeling to elucidate screening outcomes. This dissertation focuses on developing and deploying bioinformatic methods to analyze data from high-throughput genetic screens in non-conventional microbes, and identify genome-wide targets for strain engineering in an effort to improve industrial bioproduction.

Chapter 1 of this dissertation reviews the experimental and computational approaches for conducting and analyzing pooled CRISPR knockout screens in non-conventional microbes. We describe the various steps involved in the CRISPR screening workflow, ranging from sgRNA library design to interpretation of screening outcomes, while outlining the considerations and pitfalls in every step. We also discuss and compare existing software tools available for CRISPR screen analysis, sharing our insights on the suitability of different tools for different screening datasets.

The sgRNA library used to conduct pooled CRISPR screens typically consists of sgRNA with a broad spectrum of activity, with only a limited fraction of sgRNA resulting in successful genomic edits. As a result, sgRNA having low editing efficiency tend to mask the effect of genetic disruptions and consequently obscure gene hits. In Chapter 2, we address this issue by developing acCRISPR, an end-to-end pipeline to analyze data from pooled CRISPR screens and identify statistically significant genes for a given phenotype. acCRISPR uses experimental sgRNA activity profiles to correct screening outcomes by removing low-activity sgRNA based on an activity threshold. We applied acCRISPR to CRISPR-Cas9 and CRISPR-Cas12a screening datasets from the non-conventional yeast *Yarrowia lipolytica* to identify a consensus set of genes essential for growth on glucose, and previously unknown genes conferring tolerance to high salt stress conditions in industrial bioreactors, a subset of which were experimentally validated.

The experimental activity profiles used for acCRISPR analysis in Chapter 2 also serve as suitable datasets to train deep learning models and design active sgRNA based on *in silico* activity scores. Chapter 3 showcases the importance of using balanced datasets to train deep learning models for accurate prediction of high- and low-activity sgRNA. We demonstrate this by training a convolutional neural network model and a large language model on the CRISPR-Cas12a experimental activity profile from *Y. lipolytica*. Training on the original CRISPR-Cas12a data, a balanced dataset, results in accurate activity predictions for experimentally-validated high- and low-activity sgRNA, while training on imbalanced datasets obtained by removing high- or low-activity sgRNA from the Cas12a data reduces the ability to predict high- or low-activity sgRNA respectively. Moreover, we found the prediction accuracy to improve when imbalanced training sets are re-balanced by adding synthetic high- or low-activity sgRNA, while still being lower than that obtained with an inherently balanced training set.

In Chapter 4, we utilized a population genomics strategy to identify novel genetic targets responsible for the biosynthesis of phenazine compounds in the bacterium *Pseudomonas chlororaphis*. We collected and sequenced 34 Pseudomonas isolates using

Illumina and Oxford Nanopore technologies, and used the NGS reads to assemble whole genomes. We also characterized phenazine production for each isolate in two media conditions. Next, we used the whole genome assemblies along with the phenazine production data for the 34 isolates to perform a bacterial genome-wide association study (GWAS) and identified 330 significant genomic variants for improving bioproduction of various phenazines. Based on a quantitative metric, we elucidated the most influential hits for phenazine production and validated them *in vivo* in the most optimal phenazine producing strain, exemplifying the usefulness of adopting a population genomics approach to identify novel metabolic engineering targets.

Chapter 5 discusses the findings presented in this dissertation within the broader context of the field and provides future research directions.

Chapter 1: Analyzing CRISPR screens in non-conventional microbes

1.1 Abstract

The multifaceted nature of CRISPR screens has propelled advancements in the field of functional genomics. Pooled CRISPR screens involve creating programmed genetic perturbations across multiple genomic sites in a pool of host cells subjected to a challenge, empowering researchers to identify genetic causes of desirable phenotypes. These genome-wide screens have been widely used in mammalian cells to discover biological mechanisms of diseases and drive the development of targeted drugs and therapeutics. Their use in non-model organisms, especially in microbes to improve bioprocessing-relevant phenotypes, has been limited. Further compounding this issue is the lack of bioinformatic algorithms for analyzing microbial screening data with high accuracy. Here, we describe the general approach and underlying principles for conducting pooled CRISPR knockout screens in non-conventional yeasts and performing downstream analysis of the screening data, while also reviewing state-of-the-art algorithms for identification of CRISPR screening outcomes. Application of pooled CRISPR screens to non-model yeasts holds considerable potential to uncover novel metabolic engineering targets and improve industrial bioproduction.

1.2 Introduction

High-throughput CRISPR screens have become a versatile tool in enabling identification of the genetic basis of various phenotypes [1–3]. For instance, they have been used extensively in mammalian cancer cell lines to identify essential genes for survival, for facilitating targeted drug design, and in immunological studies to identify genes involved in various pathways in human immune cells [4–8]. Moreover, with the ability to target combinations of multiple genes simultaneously, CRISPR screens have made it possible to elucidate functions of poorly characterized genes via construction of gene interaction (GI) maps [9]. Genome-wide CRISPR screens have also been performed in bacteria and yeasts to unravel genetic hits influencing a diverse set of phenotypes, including those relevant to industrial bioproduction. Previous studies in model microbes – *E. coli* and *S. cerevisiae* – have identified essential genes as well as those required for conferring tolerance to biochemicals like isobutanol and furfural [10–12]. Other studies have focused on non-conventional microbes, such as the oleaginous yeast *Yarrowia lipolytica*, to identify genes essential for growth on glucose, and those important for providing tolerance to environmental stress conditions, such as low pH and high salt concentration [13,14].

CRISPR screens typically use a library of programmable single guide RNAs (sgRNAs) and a CRISPR-associated endonuclease, typically Cas9 or Cas12a, to create mutations or alter gene expression [15–17]. The most common type of CRISPR screens are knockout screens where the CRISPR-Cas system generates a double stranded break (DSB) at the genomic target site, evoking native DNA repair pathways such as non-homologous end joining (NHEJ) to create INDEL mutations that result in loss of gene function [18,19]. Besides knocking out gene function, CRISPR screens can make use of a nuclease-deactivated Cas protein to modulate transcription when fused to activator

(CRISPR activation or CRISPRa) or repressor (CRISPR interference or CRISPRi) domains. These screens can be conducted in a pooled or arrayed format. Arrayed screens physically separate predefined gene perturbations, making them malleable to amalgamation with downstream -omics profiling; but they have limited throughput and are relatively expensive [15–17]. On the contrary, pooled screens have a much higher throughput as they are devoid of physical separation between gene targets, making them more commonplace compared to arrayed screens, but require performing comparisons to a baseline for hit identification [15,16]. The customizable nature of sgRNA and the ease of inducing perturbations to gene function using CRISPR-Cas systems makes high-throughput CRISPR screens an effective tool for establishing genotype-phenotype relationships in both model and non-conventional organisms.

The CRISPR screening workflow comprises of a series of experimental and computational steps, ranging from host selection and library design to identification and biological interpretation of hits. In this review, we explore some of these steps in detail, with a focus on pooled CRISPR knockout screening in non-conventional yeasts. We begin by discussing the experimental design for conducting the screens, followed by bioinformatic processing of screening data. We also describe the general working principles behind the identification of screening outcomes, while juxtaposing the nature of yeast and mammalian cell datasets. Lastly, we review some of the existing algorithms for analyzing CRISPR screens and discuss their performance on yeast screening datasets, with a goal of assisting researchers in choosing the most appropriate tool for analyzing their data.

1.3 Functional genetic screening with pooled CRISPR libraries

A schematic representation of the experimental pipeline for performing pooled CRISPR knockout screens is depicted in **Fig. 1.1**. Since microbes exhibit different sets of desirable phenotypes, a preliminary step in CRISPR screening is the selection of an appropriate biological host for a given application. In the case of non-conventional yeasts, relevant phenotypes influencing host selection include a microorganism's natural ability to synthesize a certain bioproduct, or tolerate harsh environmental conditions that may be present in industrial bioprocesses [20]. Once a host is chosen, an sgRNA library is designed to target relevant or all protein-coding genes in the genome of the organism (although non-coding regions could also be targeted [21]).

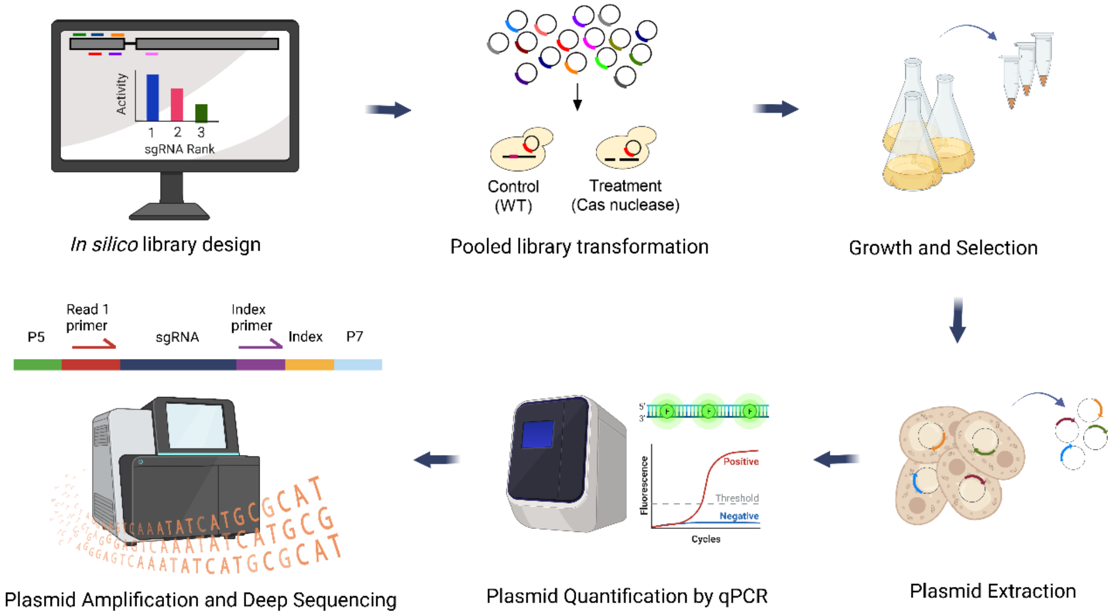


Fig. 1.1. Experimental workflow for conducting pooled CRISPR knockout screens. A library of sgRNA spanning multiple genomic sites is designed, synthesized, and cloned into a plasmid backbone. The plasmid library is transformed into control and treatment host strains, and cells are cultivated for a predetermined number of days to select for significant gene knockouts. Upon completion of the screen, plasmids are extracted from the cells, quantified by qPCR, and sequenced using NGS. Figure created with BioRender.

sgRNA library design

It is well known that guide RNAs present disparities in inducing genetic edits and that guide activity is crucial to the accuracy of screening results. Highly active guides can correlate the phenotypic variations to the appropriate genomic perturbation with high accuracy, while poorly active guides may obscure gene hits. It is thus advantageous to create a library comprising a large proportion of high-activity guides to improve hit calling. High-activity libraries can be designed by picking guides based on *in silico* activity scores estimated using activity prediction algorithms. Existing software tools such as CHOPCHOP [22], CRISPRLearner [23], DeepCpf1 [24], and DeepGuide [25], among others use one or more sequence, structural, and epigenetic features of sgRNA to predict on-target activity with endonucleases such as Cas9 and Cas12a. See [26] for an in-depth review on guide activity prediction and benchmarking across current tools. Despite most of the prediction algorithms being developed for model organisms, they have been reasonably effective in facilitating design of active sgRNA libraries in non-conventional hosts relative to an unbiased approach that is blinded to *in silico* activity scores [13,14]. Even when using activity prediction scores, it is advisable to design several guides targeting each gene (*i.e.*, a high genome coverage), ensuring the presence of an active guide per gene and sufficient statistical power for hit identification – an approach that comes with a cost of increasing downstream analytical complexity. Regardless of the design strategy, it is critical to ensure that every sgRNA in the final library is: (i) unique within the genome, so that off-target effects are minimized; (ii) does not target intronic regions of the coding sequence; (iii) sufficiently spaced from other sgRNA to improve diversity of target locations; and if possible, (iv) located within 5-65% of each coding sequence to maximize the chances of a gene knockout resulting in a non-functional protein [27,28].

Conducting the screening experiment

The designed library is synthesized as pooled single stranded oligos on a DNA microchip that are cloned into a plasmid vector, resulting in a library of plasmids. This plasmid library is amplified by transforming it into *E. coli* before isolation and subsequent transformation into the actual host cells for the screening experiment. Stable Cas expression in host cells is often accomplished through heterologous expression from a genomically integrated expression cassette. In addition to the sample (or treatment) strain, a control (or reference) strain is also needed so that changes in guide abundances at the end of the screen can be determined. In many cases, a strain devoid of Cas endonuclease but harboring the guide RNA library is used as control [1,29]. Other examples of the reference conditions include, the treatment sample immediately post transformation (day 0 of the screen before gene knockouts occur) or the untransformed library [6].

Upon transformation of the sgRNA library in the control and treatment strains, cells are allowed to proliferate until they reach confluency, and then subcultured to allow for genetic selection. At the end of the screen, the connection between genotype and phenotype is made by sequencing isolated plasmids expressing the sgRNA. That is, the fitness effect of disrupting a given gene is determined by quantifying the abundance of the sgRNA targeting the gene. To do so, the plasmid library is extracted from the treatment and control samples, the encoded guides are amplified by PCR, and the amplicon pool is sequenced using an Illumina or similar NGS platform. For accuracy of results, it is advisable to ensure sufficient depth in the sequencing run, which should be about 100 times the library size or higher for every screening replicate.

The resulting sequencing reads, the counts of which are indicative of the abundance of a given mutant in the microbial population, can be bioinformatically processed to identify genes that affect growth in the screened condition. The raw sgRNA abundances themselves cannot be used directly for accurate determination of screening outcomes, since they do not account for variability in sgRNA activity and variability in sequencing depth across samples, necessitating bioinformatic analysis to obtain significant hits from the screen.

1.4 CRISPR screen data analysis

CRISPR screen analysis pipelines typically include steps for sequencing read processing, quality control, hit identification, and investigation of screening results. These steps and a typical analysis pipeline are shown in **Fig. 1.2**, and described in detail below along with the associated bioinformatic tools available for use.

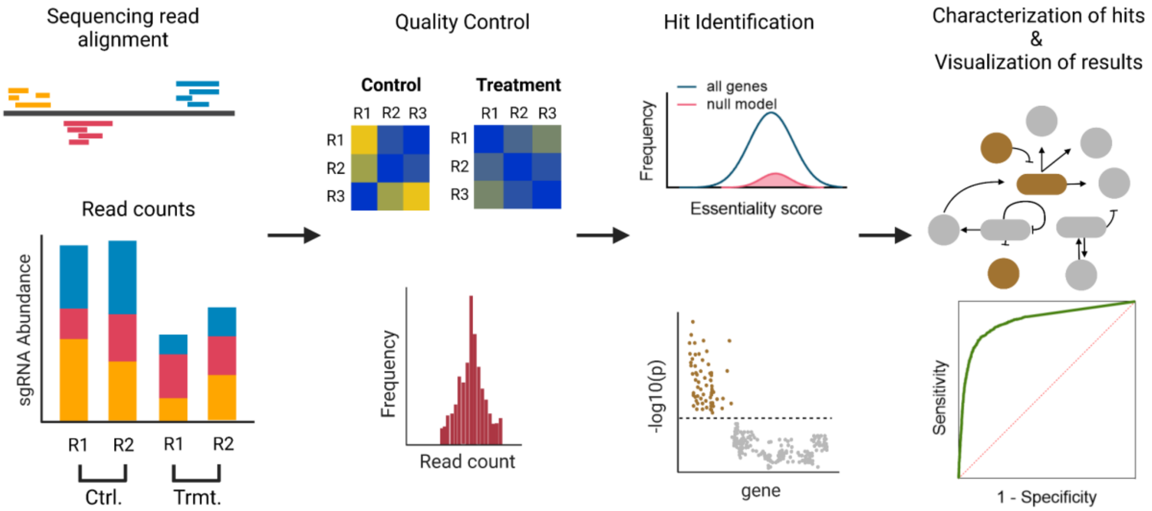


Fig. 1.2. Typical steps in pooled CRISPR screen analysis and visual depiction of results. Raw sequencing reads from the screen are processed to generate sgRNA read counts, which, upon passing quality control, are used for identifying significant genes in the dataset. The identified hits are characterized to elucidate underlying biological mechanisms and screening results are visualized by making plots. Figure created with BioRender.

Processing raw sequencing data

While analysis tools like PinAPL-Py [30] and MAGeCK [31] can accept raw sequencing data as input and process it as part of their pipeline, most other analysis packages require sgRNA abundances or log2-fold changes as input. Consequently, raw sequencing reads from the genome-wide screen need to be modified and aligned separately before analyzing them to generate screening outcomes. This can be accomplished with a number of existing bioinformatics tools and workflows. For instance, FastQC [32] allows users to perform quality control on the sequencing data based on metrics such as per base sequence quality, sequence length distribution, and overrepresented sequences, among others.

If multiple screening samples and replicates were sequenced in a single run, the reads can be demultiplexed to split the data into individual samples and replicates on the basis of sample-specific adapters. This is achieved with the help of tools such as Cutadapt [33], Ultraplex [34], or ea-utils [35]. Other tools like fastp [36] or Trimmomatic [37] could be used to trim the reads by removing the vector backbone and other miscellaneous sequences to only retain the sgRNA sequence.

In regard to mapping reads to the genome and/or the sgRNA library, available methods include BWA-MEM2 [38], Bowtie2 [39], or HISAT2 [40], which align NGS reads to a reference sequence by exact or approximate matching. Of these tools, Bowtie2 is most widely used for read mapping in CRISPR screen analysis. The read alignment information is used to compute the read count (*i.e.*, abundance) of each sgRNA across samples.

For CRISPR screens in non-model yeasts like *Yarrowia lipolytica*, the tools Cutadapt and Trimmomatic have been found to be suitable for demultiplexing and trimming sequencing reads respectively [14,25]. Similarly, a combination of Bowtie2 and naive exact matching has been shown to perform reasonably well in aligning reads, especially due to the ability of Bowtie2 to account for mismatches during alignment, that mainly stem from sequencing errors.

Quality control of the screening data

Before using the read counts for further analysis, it is essential to assess the quality of experimental data, for example, by determining pairwise replicate correlation coefficients per sample and examining the sgRNA abundance distribution in the original library. This is done to ensure authenticity of screening results and reduce spurious hit predictions. High correlation values (*e.g.*, Pearson’s coefficient > 0.7) indicate consistency between biological replicates. Upon passing this quality check, raw sgRNA read counts from control and treatment samples are provided to one or more CRISPR screen analysis methods that employ statistical approaches to identify significant genes in the screen.

Identifying screen hits

Most methods normalize the raw sgRNA abundances to account for varying sequencing depth across samples and ensure a fair comparison between controls and treatments. These normalized abundances are used directly, or converted to log2-fold change to estimate gene essentiality scores, predominantly using Bayesian principles.

The genome-wide library contains sgRNA with variable activity; failure to account for this variability could lead to inaccurate predictions of screening results. High activity sgRNA should thus have a greater influence in determining gene essentiality compared to low activity sgRNA. A common strategy to infer sgRNA activity involves screening across multiple conditions and applying probability-based approaches to make a prediction [41,42]. Alternatively, guide RNA activity can be determined experimentally by screening in an additional treatment sample containing a knockout of the native DNA repair mechanism. The activity of sgRNA can then be estimated as the log2-fold change in sgRNA abundance in the knockout-containing strain (in presence of the Cas endonuclease) to that in the control strain [13,14]. Using this approach not only improves the reliability of the activity estimate, but also avoids the need to screen across multiple conditions (substantially reducing the size of the experiment), although knockout of DNA repair may not always be viable for all organisms.

Once essentiality scores have been computed, a statistical test for every gene to determine whether it belongs to a 'null' population of scores (*i.e.*, population of essentiality scores of non-essential genes) is typically conducted, thus resulting in a p-value for the essentiality of each gene. The p-value is further corrected for multiple comparisons (typically using FDR [43]) and genes having corrected p-value lower than a predetermined threshold are deemed as significant hits or essential genes.

Selecting a null model for significance testing

A suitable choice for the 'null' population depends on the host organism used to generate the screening data. Ideally, the null population is representative of the behavior of non-essential genes in the screen. For mammalian cells, the non-essential gene population overlaps well with the population of negative control sgRNA, and as a result, the negative control population serves as a suitable null model. On the other hand, non-conventional yeast datasets, in our experience, have a non-essential gene population that is relatively distant from the population of negative control sgRNA [14]. Using negative controls to create the null population would thereby result in a large number of false essentiality predictions, prompting the use of putative non-essential genes to create the null population.

While negative control sgRNA make no knockouts in the genome, knockouts produced by targeting sgRNA result in growth defects and a corresponding drop in the targeting sgRNA abundance compared to the control sample. The proximity between the negative control sgRNA and non-essential gene populations thus depends on the ability of host cells to stem these growth defects. Non-conventional yeasts lack this ability to suppress growth defects, likely due to the absence of multiple gene copies and alternate splicing mechanism. This is in contrast to the case of mammalian cells, which exhibit polyploidy and undergo alternate splicing of genes, presumably suppressing growth defects and causing the non-essential gene population to overlap well with the negative control population.

Investigation of screening results

After identifying significant genes from a screen, the next step is to elucidate their biological importance. Databases such as UnitProt [44] and Pfam [45] can be used to investigate protein functions of known genes. In addition, analyses like gene ontology (GO)-enrichment test [46] and GSEA [47] could be performed to identify biological pathways relevant to the significant hits. Since non-model organisms have a considerable number of unannotated genes, these could be investigated by performing BLAST [48] against proteomes of model organisms, or more rigorously by experimentation. Finally, screening results can be visualized, for example, by plotting log2-fold changes of sgRNA targeting significant genes against a backdrop of those of the entire library. Moreover, if a gold standard set of essential genes is available, receiver operator characteristic (ROC) plots or precision-recall (PR) plots can be constructed and area under the curve can be calculated to determine accuracy of the predictions.

1.5 Software packages for CRISPR screen analysis

Here we introduce and describe the most commonly used software packages for analyzing pooled CRISPR screens. A comparison of the tools based on some common features is provided in **Table 1.1**.

Table 1.1. Comparison of software packages for analysis of pooled CRISPR screens.

Method	Implement-ation	Quality control	Expt. sgRNA efficiency	Multiple screens	Applicable to CRISPRa/i	Ref.
MAGeCK-VISPR	Python	Yes	No	Yes	No	[41]
CRISPhieRmix	R,C++	No	No	No	Yes	[49]
JACKS	Python	No	No	Yes	No	[42]
ACE	R	No	No	Yes	No	[53]
BAGEL2	Python	Yes	No	No	No	[51]
acCRISPR	Python	No	Yes	No	No	[14]

MAGeCK-VISPR

MAGeCK-VISPR is an end-to-end workflow for quality control, analysis, and visualization of CRISPR screens [41]. The analysis is carried out by an expectation-maximization algorithm that takes raw sgRNA counts from multiple screening conditions as input, and uses them to iteratively compute gene essentiality across conditions and sgRNA activity. Read counts are modeled as a negative binomial distribution and a generalized linear model is used to deconvolute gene effects from multiple screens. Although shown to be robust in making predictions for mammalian cancer cell lines, the method inaccurately estimates sgRNA activity for datasets from non-conventional yeasts, which leads to erroneous predictions for gene essentiality [14].

CRISPhieRmix

Originally designed to analyze CRISPRa and CRISPRi screens, CRISPhieRmix can also be applied to knockout screens [49]. The method requires log2-fold changes of sgRNA as input and fits that data to a hierarchical mixture distribution, constituting a broad tailed null distribution (to account for asymmetry in the screening data) and an alternative distribution. This model is used to compute and return the posterior probability of belonging to the alternative distribution for each gene, marginalized over all possible mixture distributions of sgRNA targeting essential genes. Since CRISPhieRmix uses the negative control population to form the null distribution, it performs well on screening data from human cancer cells, but has been found to result in an excessive number of false positives for screening datasets in the yeast *Yarrowia lipolytica* [14].

JACKS

JACKS is a Bayesian method that processes data from multiple screens simultaneously to improve modeling of sgRNA activity and hence, estimation of condition-dependent gene essentiality [42]. The method starts out by assuming Gaussian priors for gene essentiality scores and sgRNA efficacies. It further uses raw sgRNA counts as input to compute log2-fold changes that constitute the likelihood function. The final values of sgRNA activity and gene essentiality per condition are inferred from their respective posteriors, determined using variational inference. Like MAGeCK-MLE, JACKS effectively identifies essential genes in human datasets, but has been shown to fall short of correctly classifying essential genes in non-conventional yeasts like

Yarrowia lipolytica, mainly due to its inability to make accurate sgRNA activity inferences [14].

BAGEL2

Developed as an updated version of BAGEL [50], this method uses information from gold-standard sets of essential and non-essential genes to infer essentiality of every gene in the screening dataset, via calculation of a Bayes factor corrected for off-target effects [51]. BAGEL2 accounts for copy number effect using the tool CRISPRcleanR [52]. Additionally, it determines the quality of each screening replicate by computing a quality score based on log-fold change of sgRNA targeting reference essential and non-essential genes. Since gold-standard sets of essential and non-essential genes may not always be available, as is the case with most non-model organisms, this method may have limited cross-species applicability at present.

ACE

ACE is a probabilistic method with the ability to predict differential gene essentiality between samples, in addition to absolute essentiality [53]. The method does this using sgRNA abundance in the untransformed library, along with initial and final abundances from each screening sample, which are all modeled as Poisson distributions and help define the likelihood function. Knockout efficiency of sgRNA is computed using a logistic regression model, assuming that it depends on the GC content of each guide sequence. Finally, ACE estimates gene essentiality and the logistic regression coefficients iteratively using maximum-likelihood estimation and determines gene significance from separate likelihood ratio tests for absolute and differential essentiality. Thus far, this analysis package has primarily been used to successfully identify gene essentiality in mammalian cancer cell lines [53].

acCRISPR

acCRISPR is an activity-correction method that improves CRISPR screening outcomes by optimizing sgRNA library activity [14]. The method uses experimental sgRNA efficiency profiles, obtained by knocking out the dominant host DNA repair mechanism (such as non-homologous end joining by deletion of *ku70* gene), to remove low activity sgRNA from the analysis and correct screening outcomes based on an activity threshold, by calculating an ac-coefficient (given as the product of sgRNA activity threshold and library coverage). In absence of experimental activity values, acCRISPR can utilize predicted activity scores for the library, if available. Gene essentiality is determined by testing against a null distribution, created using sgRNA targeting putative non-essential genes, which makes acCRISPR a suitable method for analyzing screens in non-model yeasts. This method has been shown to accurately call essential genes and genes important for environmental stress tolerance in the oleaginous yeast *Yarrowia lipolytica* [14].

1.6 Conclusions and Perspectives

Pooled CRISPR screens have shown great promise in facilitating biological discovery by enabling identification of genetic signatures for known and novel phenotypes. Although CRISPR screens have been extensively used in mammalian cells to investigate disease mechanisms, their application to non-conventional microbes for improving metabolic engineering-relevant phenotypes has been limited so far. This review describes the experimental and computational steps involved in conducting and analyzing CRISPR knockout screens, with a focus on approaches and methods that have been successfully deployed in non-conventional yeasts. While the integration of these steps makes for an end-to-end workflow, there are several considerations that one needs to be mindful of in the entire process.

The ability of the sgRNA library to produce genetic knockouts, for instance, plays a pivotal role in determining the effectiveness of a screen. Accordingly, libraries should be formulated to include as many high-activity guides as possible. This could be achieved in part by using activity scores obtained from sgRNA activity prediction tools to inform library design. In the absence of accurate activity predictions, as is often the case with most non-model organisms, it is advisable to create a library having high genome coverage to ensure sufficient statistical power in evaluating screening outcomes.

Another key aspect in improving screen design and analysis is the successful delineation of sgRNA activity profiles in the context of the screen itself. While predicted activity scores may be readily available, sgRNA efficiencies are susceptible to variation in the screening environment, warranting this extra measurement. Such activity profiles can be derived experimentally, or by modeling single or multiple screens. This additional data can be leveraged to diminish the influence of low-activity sgRNA in estimating gene effects, thereby enhancing the accuracy of hit identification. Other considerations for optimizing experimental design of the screen include using an adequate number of biological replicates, ensuring high library representation at the start of the screen, and sequencing the library at a sufficient depth.

Overall, CRISPR knockout screening in non-conventional microbes is an evolving tool that could be harnessed to investigate biological mechanisms and thus decode the genetics of the host organism. In addition to knockout screening, future studies should focus on knockdown and activation screens (CRISPRi and CRISPRa, respectively), promoting discovery of gene function and establishment of novel genotype-phenotype relationships. These biological findings would further improve host genetic engineering, drive enhancement of desirable phenotypes, and consequently improve the feasibility of industrial bioprocesses.

1.7 References

1. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515–1526 (2015).

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

3. Doench, J. G. Am I ready for CRISPR? A user's guide to genetic screens. *Nat. Rev. Genet.* **19**, 67–80 (2018).

4. Tzelepis, K. *et al.* A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia. *Cell Rep.* **17**, 1193–1205 (2016).

5. Aguirre, A. J. *et al.* Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov.* **6**, 914–929 (2016).

6. Meyers, R. M. *et al.* Computational correction of copy number effect improves specificity of CRISPR–Cas9 essentiality screens in cancer cells. *Nat. Genet.* **49**, 1779–1784 (2017).

7. Shifrut, E. *et al.* Genome-wide CRISPR Screens in Primary Human T Cells Reveal Key Regulators of Immune Function. *Cell* **175**, 1958–1971.e15 (2018).

8. Parnas, O. *et al.* A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell* **162**, 675–686 (2015).

9. Horlbeck, M. A. *et al.* Mapping the Genetic Landscape of Human Cells. *Cell* **174**, 953–967.e22 (2018).

10. Rousset, F. *et al.* Genome-wide CRISPR-dCas9 screens in E. coli identify essential genes and phage host factors. *PLoS Genet.* **14**, e1007749 (2018).

11. Wang, T. *et al.* Pooled CRISPR interference screening enables genome-scale functional genomics study in bacteria with superior performance. *Nat. Commun.* **9**, 2475 (2018).

12. Bao, Z. *et al.* Genome-scale engineering of Saccharomyces cerevisiae with single-nucleotide precision. *Nat. Biotechnol.* **36**, 505–508 (2018).

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

14. Ramesh, A. *et al.* acCRISPR: An activity-correction method for improving the accuracy of CRISPR screens. *bioRxiv* 2022.07.12.499789 (2022) doi:10.1101/2022.07.12.499789.

15. Bodapati, S., Daley, T. P., Lin, X., Zou, J. & Qi, L. S. A benchmark of algorithms for the analysis of pooled CRISPR screens. *Genome Biol.* **21**, 62 (2020).

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

17. Kampmann, M. CRISPRi and CRISPRa Screens in Mammalian Cells for Precision Biology and Medicine. *ACS Chem. Biol.* **13**, 406–416 (2018).

18. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).

19. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).

20. 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).

21. Shukla, A. & Huangfu, D. Decoding the noncoding genome via large-scale CRISPR screens. *Curr. Opin. Genet. Dev.* **52**, 70–76 (2018).

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

23. Dimauro, G. *et al.* CRISPRLearner: A Deep Learning-Based System to Predict CRISPR/Cas9 sgRNA On-Target Cleavage Efficiency. *Electronics* **8**, 1478 (2019).

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

25. 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).

26. Konstantakos, V., Nentidis, A., Krithara, A. & Paliouras, G. CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Res.* **50**, 3616–3637 (2022).

27. 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).

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

29. Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* **32**, 267–273 (2014).

30. Spahn, P. N. *et al.* PinAPL-Py: A comprehensive web-application for the analysis of CRISPR/Cas9 screens. *Sci. Rep.* **7**, 15854 (2017).

31. Wang, B. *et al.* Integrative analysis of pooled CRISPR genetic screens using MAGeCKFlute. *Nat. Protoc.* **14**, 756–780 (2019).

32. Andrews, S. & Others. FastQC: a quality control tool for high throughput sequence data. Preprint at (2010).

33. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

34. Wilkins, O. G., Capitanchik, C., Luscombe, N. M. & Ule, J. Ultraplex: A rapid, flexible, all-in-one fastq demultiplexer. *Wellcome Open Res* **6**, 141 (2021).

35. Aronesty, E. ea-utils: Command-line tools for processing biological sequencing data. Preprint at (2011).

36. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

37. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

38. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324 (2019).

39. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

40. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

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

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

43. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).

44. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–12 (2015).

45. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405–420 (1997).

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

47. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).

48. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

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

50. Hart, T. & Moffat, J. BAGEL: a computational framework for identifying essential genes from pooled library screens. *BMC Bioinformatics* **17**, 164 (2016).

51. Kim, E. & Hart, T. Improved analysis of CRISPR fitness screens and reduced off-target effects with the BAGEL2 gene essentiality classifier. *Genome Med.* **13**, 2 (2021).

52. Iorio, F. *et al.* Unsupervised correction of gene-independent cell responses to CRISPR-Cas9 targeting. *BMC Genomics* **19**, 604 (2018).

53. Hutton, E. R., Vakoc, C. R. & Siepel, A. ACE: a probabilistic model for characterizing gene-level essentiality in CRISPR screens. *Genome Biol.* **22**, 278 (2021).

Chapter 2: acCRISPR: An activity-correction method for improving the accuracy of CRISPR screens

2.1 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. Here, 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 via calculation of an optimization metric, thus determining the fitness effect of disrupted genes. CRISPR-Cas9 and -Cas12a screens were carried out in the non-conventional oleaginous yeast *Yarrowia lipolytica* to determine 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 screens quantifying relative cellular fitness under high salt conditions to identify known and novel genes that were related to salt tolerance. 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.

2.2 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. 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 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 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 high salt tolerance screens. We independently validate acCRISPR predictions by measuring growth of individual disruptions of a subset of essential genes and tolerance genes in conditions akin to those of the original genome-wide screens. We also evaluate the performance of acCRISPR with computational predictions of guide activity rather than experimentally determined values. 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.

2.3 Results

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. 2.1a**). 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 2.1 and 2.2**.

The first analytical step of acCRISPR is to convert raw guide abundance values into CS and FS profiles (**Fig. 2.1b, Supplementary File 2.3**). 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).

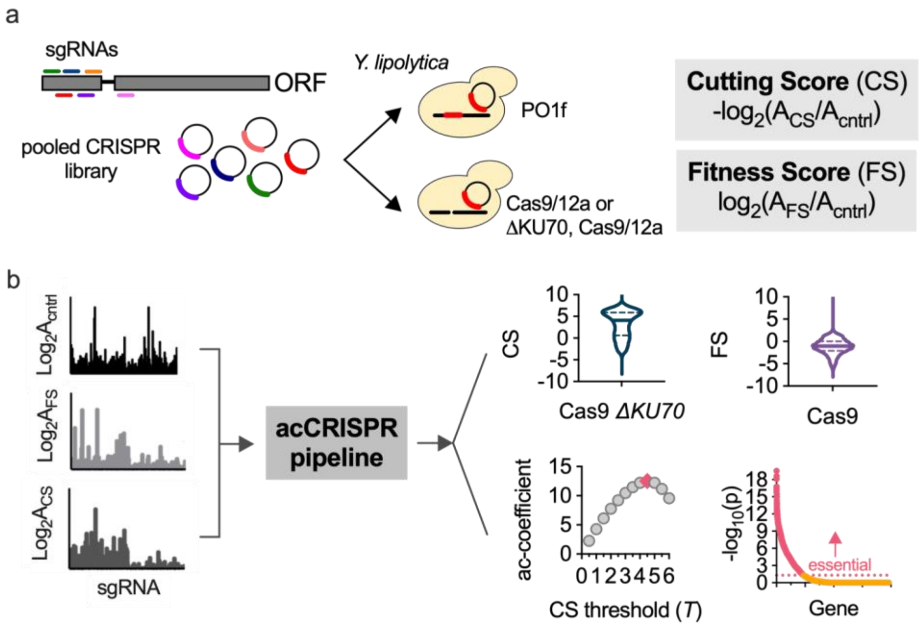


Figure 2.1. acCRISPR analysis of CRISPR-Cas screens. (**a**) Growth screens in *Y. lipolytica* were conducted with pooled libraries of single guide RNAs (sgRNAs) (6- and 8-fold coverage of >98.5% of CDSs, for Cas9 and Cas12a respectively). A guide's cutting score (CS) is equal to the $-log_2$ fold-change 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.

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. 2.2**. Similar analyses of the CRISPR-Cas12a screens are shown in **Fig. S2.1**.

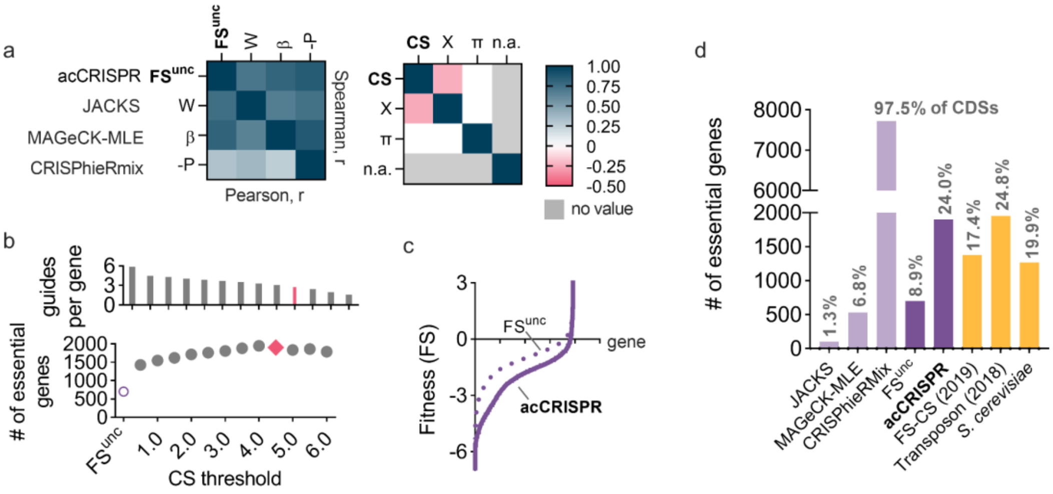


Figure 2.2. acCRISPR analysis of CRISPR-Cas9 screens defines a high confidence set of essential genes. (**a**) Heat maps showing Pearson (below diagonal) and Spearman (above diagonal) correlation coefficients for comparison of gene fitness effects (uncorrected FS (FS^{unc}), W, β, and -P; left) and sgRNA 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[9] 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. 2.2a**). 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. 2.1b**. 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. 2.2b**. 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 (**Fig. S2.1**).

The optimized acCRISPR analysis of the Cas9 screen identified 1903 essential genes (see **Supplementary File 2.4**), 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 2.4**) 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).

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 2.5 and 2.6** 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 GO-Slim terms are shown in **Fig. 2.3a**.

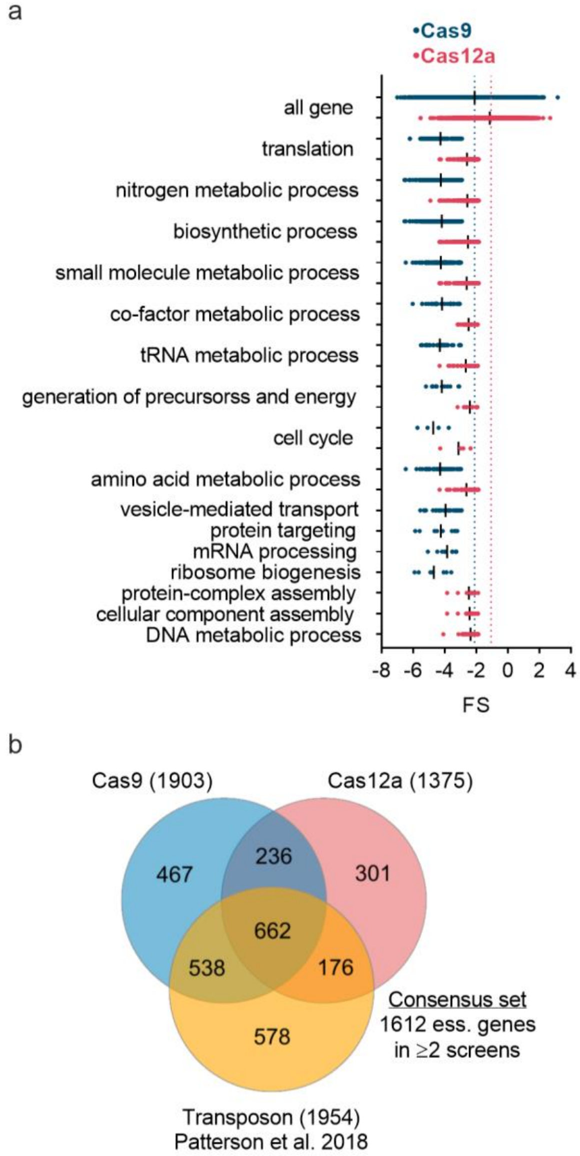


Figure 2.3. Defining a set of consensus essential genes in *Y. lipolytica*. (**a**) Enriched GO-Slim biological process terms for Cas9 and Cas12a essential gene sets and FS distribution of essential genes associated with each GO-Slim term. Enriched terms were determined using a hypergeometric test (FDR-corrected, $p < 0.05$). 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.

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. 2.3b** and

Supplementary File 2.7). 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 2.8**). To verify the essentiality of genes in the consensus set, we tested 15 essential genes from this set and 5 non-essential genes (*i.e.*, genes non-essential in all 3 screens) using the CRISPR-Cas9 system and measured their abundance in glucose after 2 days of growth (**Fig. S2.2**; see Methods for details on the experimental procedure). Of the 15 essential genes tested, 12 were called as essential in all three screens, while 3 others were called as essential in the Cas9 and Cas12a screens, but not in the transposon screen. As expected, cells containing essential gene knockouts showed no or minimum growth throughout the validation experiment, whereas disruptions of non-essential genes exhibited substantial growth over the same time period. One-tailed t-test indicates that the growth of non-essential gene knockouts is significantly higher ($p < 0.0001$) than that of the essential gene knockouts. 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 (**Fig. S2.3**).

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. 2.4** and **Supplementary File 2.9**). 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 (**Fig. S2.4 and Supplementary File 2.10**). Using the predicted activity scores, we implemented acCRISPR to compute the maximum ac-coefficient (**Table S2.1**) and determined a set of predicted essential genes. The consensus set identified in **Fig. 2.3** 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. 2.4** and **Fig. S2.4**). 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 **Fig. S2.4**; the Cas12a screen captured 66.7% of the consensus set with 78.1% precision.

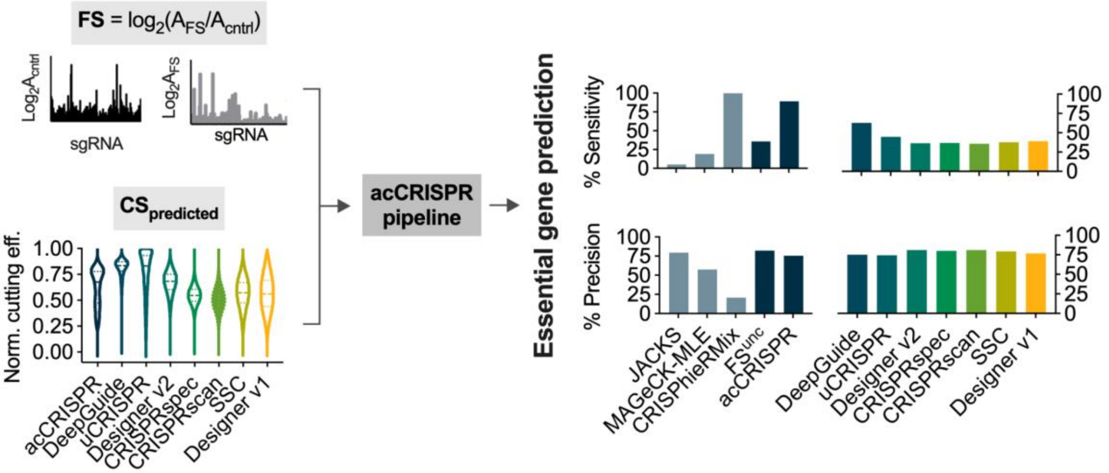


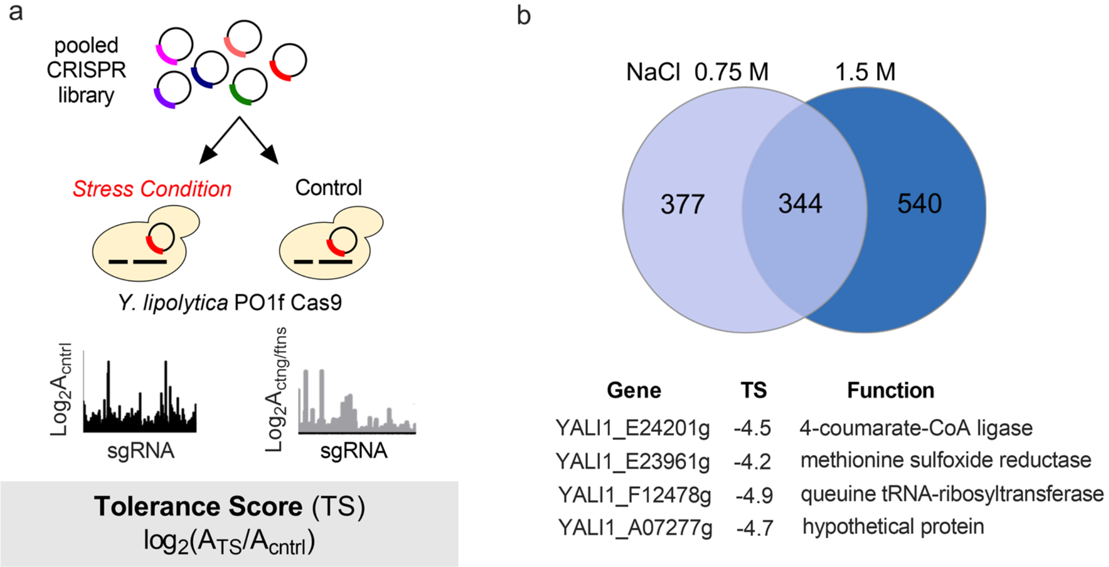
Figure 2.4. Performance of acCRISPR using predicted sgRNA activity profiles in *Y. lipolytica*. Raw sgRNA counts from control and treatment strains used for fitness score calculations were provided as input to acCRISPR along with sgRNA activity scores from a range of guide prediction tools (DeepGuide [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.

acCRISPR identifies biologically insightful hits related to salt tolerance

To further demonstrate the utility of acCRISPR, we conducted high salt tolerance screens from which we identified genetic hits that produced significant effects on cell fitness. Tolerance to high salinity is an industrially beneficial trait that can reduce costs associated with process sterilization and enable growth in lower-cost water sources (*e.g.*, seawater or wastewater) [31]. The CRISPR-Cas9 strain was grown in the presence and absence of two different levels of salt concentration ([NaCl] of 0.75 and 1.5 M) and acCRISPR was used to identify significant hits for each salt stress condition. As a control, the Cas9-containing strain was grown under standard growth conditions (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 (*i.e.*, in the presence of salt) to that grown under control conditions (**Fig. 2.5a**). A low TS indicated that gene disruption conferred a growth disadvantage under the applied stress (see **Fig. S2.5** for corrected TS profiles in tolerance screens conducted at 0.75 M and 1.5 M NaCl).

acCRISPR analysis of the salt tolerance screens (**Fig. S2.6**) identified 721 and 884 gene hits in 0.75 M and 1.5 M NaCl respectively (**Supplementary File 2.11**). The two screening conditions were found to share 344 significant genes in common (**Fig. 2.5b**). Similar to the essential gene screening outcomes, we sought to validate a subset of the gene hits (see Methods for experimental details). The validation set included four genes: YALI1_E24201g ($TS_{1.5M\ NaCl}$ = -4.5), YALI1_E23961g ($TS_{1.5M\ NaCl}$ = -4.2), YALI1_F12478g ($TS_{1.5M\ NaCl}$ = -4.9), and YALI1_A07277g ($TS_{1.5M\ NaCl}$ = -4.7; significant only in 1.5 M NaCl). YALI1_E24201g and YALI1_E23961g were selected for validation because homologs of these genes are known to affect salt tolerance in other species. The GO-term of YALI1_E24201g suggests this gene encodes for 4-coumarate-CoA ligase, which has been shown to enhance abiotic stress tolerance, including salt tolerance, in various plant species [32–34]. YALI1_E23961g is homologous to methionine sulfoxide reductase (*MXR1*) in *S. cerevisiae* and has been shown to improve resistance to oxidative stress in *S. cerevisiae* [35]. The other two gene hits selected for validation, YALI1_F12478g (a queuine tRNA-ribosyltransferase) and YALI1_A07277g (a hypothetical protein), have no known connection to stress tolerance. In all four cases, gene disruption in individual experiments that mimicked the screening conditions resulted in significantly lower ($p < 0.01$) growth than the disruption of a gene with a higher TS value that was not called as significant by acCRISPR, thus validating the called hits (**Fig. S2.7**).

Overall, the results reported here support the validity of our acCRISPR analysis in identifying novel gene hits related to salt 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.



Gene	TS	Function
YALI1_E24201g	-4.5	4-coumarate-CoA ligase
YALI1_E23961g	-4.2	methionine sulfoxide reductase
YALI1_F12478g	-4.9	queuine tRNA-ribosyltransferase
YALI1_A07277g	-4.7	hypothetical protein

Figure 2.5. acCRISPR analysis of salt 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 tolerance screens. Venn diagram (top) shows the overlap of gene hits identified in the salt (0.75 M and 1.5 M NaCl) screens. Selected hits are shown (bottom), including the gene ID, the TS value from the 1.5 M NaCl condition, and putative gene function.

2.4 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, resolving 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* (**Fig. S2.8** and **see refs.** [18,36]). 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 2.3**). 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. 2.4**). 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 [8,16,37]. 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, a subset of which were independently validated experimentally (**Fig. 2.3** and **Fig. S2.2**). 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 [38,39]; 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 2.8**), while those unique to each method have no enriched functions (**Supplementary File 2.12**).

With respect to the high salt tolerance screens, acCRISPR analysis also helps to advance our understanding of *Yarrowia* genetics by identifying high confidence hits with significantly decreased cell fitness, a subset of which were independently validated. This information promises to guide future strain engineering seeking to improve production host tolerance to harsh environmental conditions.

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 calling loss of fitness hits for growth under high salt conditions. 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.

2.5 Methods

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 2.13** for replicate correlations of sgRNA abundance in control and treatment samples for Cas9 and Cas12a screens). Let us call C_1 and T_1 the control and treatment samples, respectively, for determining cutting scores. The cutting score CS_i of sgRNA i is defined as follows

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

where $\underline{x}_{C_1,i}$ and $\underline{x}_{T_1,i}$ indicate the total normalized read counts of sgRNA i in samples C_1 and T_1, 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

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

where $\underline{x}_{C_2,i}$ and $\underline{x}_{T_2,i}$ 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 m_g 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]. In addition, studies in mammalian cells have identified ~20% or fewer genes as essential for survival of various cell lines of interest [40–43]. 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.

acCRISPR also has the ability to analyze CRISPR screening data to identify both loss- and gain-of-function hits (LOF and GOF). In this case, the fraction of genes directly related to the phenotype is typically less than the number of essential genes. Thus, we assume that 95% of genes in the screening dataset (*i.e.*, FS 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.

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 **Table S2.1**.

For the salt tolerance datasets, raw sgRNA counts from the control and treatment samples were used to calculate TS for each sgRNA (in the same manner as FS calculation) in the specific screening condition. These sgRNA TS values were used as input to acCRISPR in conjunction with the already calculated CS values from the essential gene analysis. Before implementing acCRISPR, sgRNA having very low normalized abundance (< 2.5% of the mean normalized abundance) in the control sample for TS calculation were discarded from the library. Significant genes from acCRISPR were then determined by performing a one-tailed test of significance. In all cases, genes having FDR-corrected p-value less than 0.05 were considered as significant.

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 that of 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.

Microbial strains and culturing

All strains used in this work are presented in **Table S2.2**. 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 [9,44]. 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 0.75 M and 1.5 M sodium chloride. The desired salinity was achieved by the addition of an appropriate quantity of autoclaved 5 M sodium chloride stock solution.

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.

Plasmid construction

All plasmids and primers used in this work are listed in **Tables S2.3 and S2.4**. 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 [44]. 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 BioLabs (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 construed 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).

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][45]) 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].

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 pLeumock-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.

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 **Table S2.5**.

Transformation for the 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 RPM shaking, 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.

Y. lipolytica salt tolerance screens

CRISPR-Cas9 growth screens with high salinity were conducted in synthetic defined media deficient in leucine. Media were prepared with two different salt 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 three different media: SD-leu, SD-leu (0.75 M NaCl), and SD-leu (1.5 M NaCl). 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 salt tolerance screens to select for genetic perturbations that conferred a growth disadvantage only under the stressed condition.

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 (**Table S2.6**). Schematics of the amplicons from the Cas9 and Cas12a screens submitted for NGS are depicted in **Fig. S2.9**. 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 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 (**Table S2.7**). 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.

Generating sgRNA read counts from raw reads

Next-generation sequencing raw fastq files were processed using the Galaxy platform [46]. 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 Cas9 screens are provided in **Tables S2.8 and S2.9** respectively. MATLAB scripts are provided as part of the GitHub link found below in the “Code availability” section. **Supplementary File 2.14** provides further information correlating the NCBI SRA file names to the information needed for demultiplexing the readsets. Analysis of raw 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 2.15** for raw sgRNA counts of the untransformed Cas9 and Cas12a libraries).

Gene ontology enrichment analysis

GO annotations for the CLIB89 reference genome of *Y. lipolytica* [47] were obtained from MycoCosm (mycocosm.jgi.doe.gov). GO analysis for the essential gene sets was performed using the Galaxy platform [46]. 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.

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.

Experimental validation of essential genes and salt tolerance genes

Selected hits from the essential gene and salt tolerance screens were validated by performing single gene knockouts using CRISPR-Cas9 genome editing and measuring the growth of these knockouts. Gene knockouts were made by using high-activity sgRNAs (*i.e.*, sgRNA with cutting scores greater than 5.0; see **Table S2.10** for a complete list). For construction of sgRNA expression vector, pCas9yl-GW was digested with AvrII, similar to the construction of sgRNA library plasmids. Primers for sgRNA cloning were obtained from Integrated DNA Technology (IDT). Each primer contained 20 bp of homology flanking either side of a 20 bp target sequence. A mixture of two primers was placed in a thermocycler to anneal the oligos together and create double stranded DNA. Next, the annealed oligonucleotide was inserted by HiFi DNA Assembly (New England BioLabs, NEB) into a linearized pCas9yl-GW vector. Successful cloning of the sgRNA fragment was confirmed by Sanger sequencing.

Cells containing integrated Cas9 were grown in YPD before being subjected to transformation of plasmid containing an sgRNA. All transformants were then inoculated in 17 x 100 mm round-bottomed polystyrene tubes containing 3 mL of SD-Leu media and allowed to grow for 16 hours at 30 °C and 200 rpm shaking. Cells were then subcultured in 2 mL of fresh media with a starting OD_{600} of 0.025. After 2 days of growth, cell density was determined by measuring OD_{600} using a Nanodrop 2000c (Fisher Scientific) and a 1 cm pathlength cuvette. In the case of essential genes, a culture containing cells with an empty vector was used as a positive control, while the wildtype strain containing no plasmid was used as a negative control. Two biological replicates were performed for each sample.

Validation of salt tolerance genes was performed using high salinity media (SD-Leu containing 1.5 M NaCl). Cas9 expressing cells were transformed with plasmid containing sgRNA and transformants were grown in SD-Leu for 16 hours. This was followed by inoculation in 2 mL of high salinity media to an initial OD_{600} of 0.025. Inoculation in SD-Leu devoid of salt was used as a reference condition. After 4 days of growth in the presence and absence of salt stress, cell density was determined by measuring the OD_{600}. Sample containing cells with an empty plasmid was used as a positive control. Two biological replicates were performed for each sample.

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 [48]. Similarly, DeepCpf1 scores were obtained using DeepCpf1 [30].

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:

$$\%\,Sensitivity = \left(\frac{No.\,of\ predicted\ essential\ genes\ overlapping\ with\ the\ consensus\ set}{Size\ of\ the\ consensus\ set}\right) * 100$$

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

$$\%\,Precision = \left(\frac{No.\,of\ predicted\ essential\ genes\ overlapping\ with\ the\ consensus\ set}{Total\ no.\,of\ predicted\ essential\ genes}\right) * 100$$

2.6 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.

2.7 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. A permanent repository of the software has been created and archived to Zenodo (https://doi.org/10.5281/zenodo.7847623 [49]).

2.8 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. Wang, C.-H. *et al.* Characterization and Functional Analysis of 4-Coumarate:CoA Ligase Genes in Mul-berry. *PLoS One* **11**, e0155814 (2016).

33. Chen, X., Su, W., Zhang, H., Zhan, Y. & Zeng, F. Fraxinus mandshurica 4-coumarate-CoA ligase 2 enhances drought and osmotic stress tolerance of tobacco by increasing coniferyl alcohol content. *Plant Physiol. Biochem.* **155**, 697–708 (2020).

34. Song, Z. *et al.* Melatonin enhances stress tolerance in pigeon pea by promoting flavonoid enrichment, particularly luteolin in response to salt stress. *J. Exp. Bot.* **73**, 5992–6008 (2022).

35. Moskovitz, J. *et al.* Overexpression of peptide-methionine sulfoxide reductase in Saccharomyces cerevisiae and human T cells provides them with high resistance to oxidative stress. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14071–14075 (1998).

36. 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).

37. 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).

38. 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).

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

40. Yilmaz, A., Peretz, M., Aharony, A., Sagi, I. & Benvenisty, N. Defining essential genes for human pluripotent stem cells by CRISPR–Cas9 screening in haploid cells. *Nat. Cell Biol.* **20**, 610–619 (2018).

41. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).

42. Blomen, V. A. *et al.* Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092–1096 (2015).

43. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**, 1515–1526 (2015).

44. 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).

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

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

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

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

49. Trivedi, V., Ramesh, A. & Wheeldon, I. *ianwheeldon/acCRISPR: acCRISPR first release*. (2023). doi:10.5281/zenodo.7847623.

2.9 Supplementary Information

Supplementary Figures

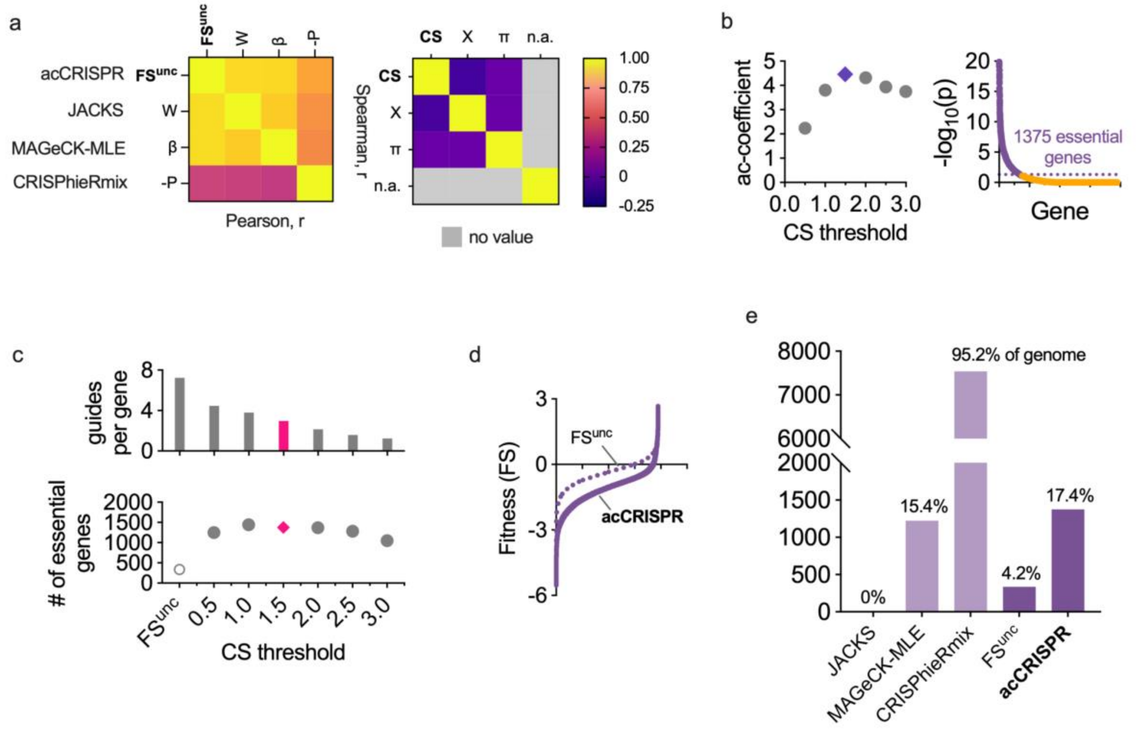


Fig. S2.1. acCRISPR analysis of Cas12a growth screens in *Yarrowia lipolytica*. (**a**) Heat-maps showing Pearson (below diagonal) and Spearman (above diagonal) coefficients of fitness effects (uncorrected FS (FS^{unc}), W, β & -P; left) and sgRNA cutting efficiencies (CS, X and π; right) and from acCRISPR 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.

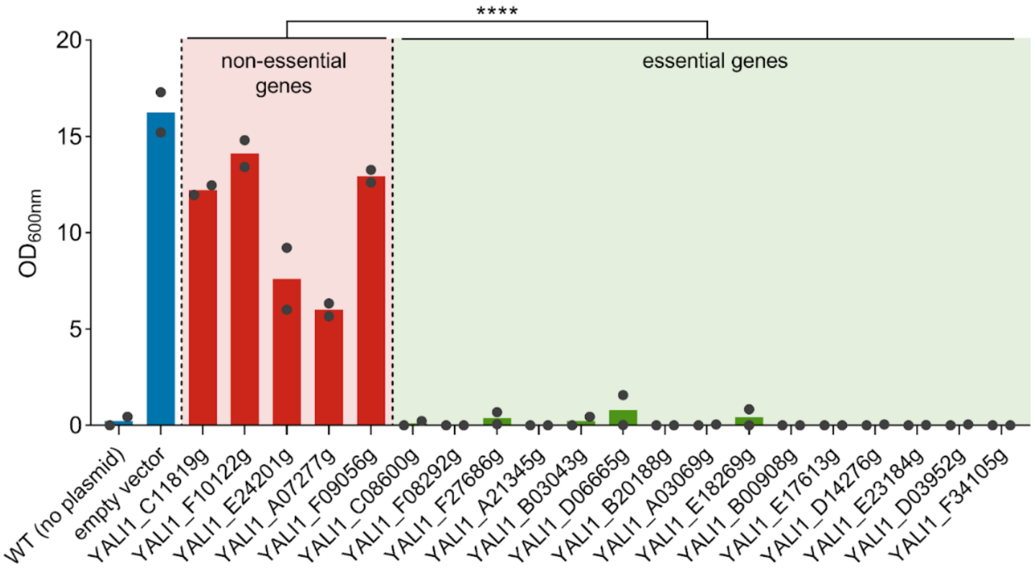


Fig. S2.2. Experimental validation of CRISPR-Cas9 and CRISPR-Cas12a essential and non-essential genes from acCRISPR analysis. Final OD of cells containing single knockouts of 5 non-essential genes (red bars) and 15 essential genes (green bars) from the consensus set. Of the 15 selected for validation, 12 were called as essential genes in all 3 screens (Cas9, Cas12a and transposon [4]). The other three genes, YALI1_B03043g, YALI1_E18269g and YALI1_F34105g, were called as essential only in the Cas9 and Cas12a screens. Cells were grown in SD-Leu for 16 hrs post sgRNA transformation, followed by subculturing in fresh media and growth for another 2 days before measuring final OD. An empty vector control (blue bar) was used to show growth in absence of any knockout. The PO1f strain containing no plasmid (indicated as WT; leftmost bar) was used as no growth, negative control. Each bar represents mean of two biological replicates (n = 2), and data points represent OD of each individual replicate in the respective sample (****$p < 0.0001$; one-tailed unpaired t-test).

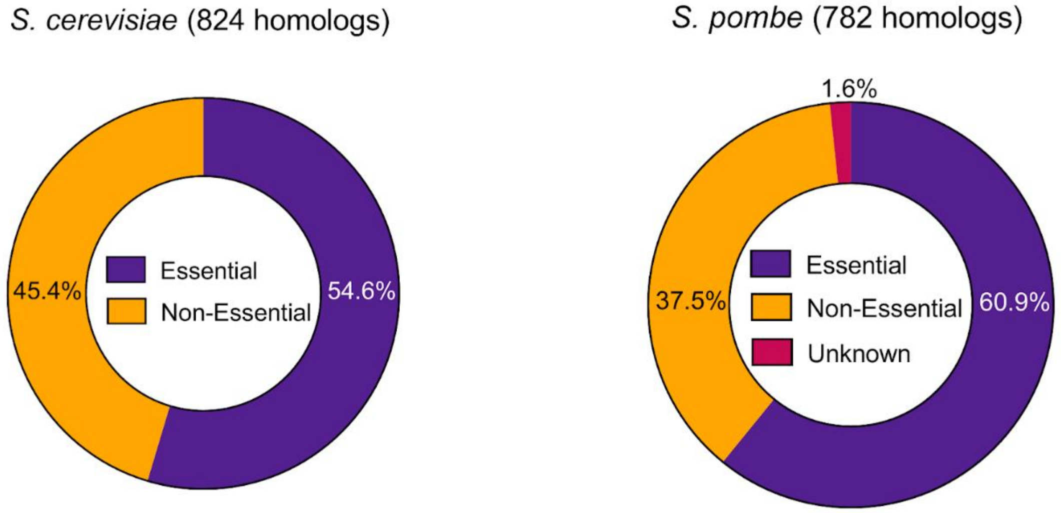


Fig. S2.3. Essential gene comparison to *S. cerevisiae* and *S. pombe*. Pie charts indicating the percentage of homologs in the *Y. lipolytica* consensus set that are essential, non-essential and have unknown essentiality in *S. cerevisiae* (824 homologs) and *S. pombe* (782 homologs).

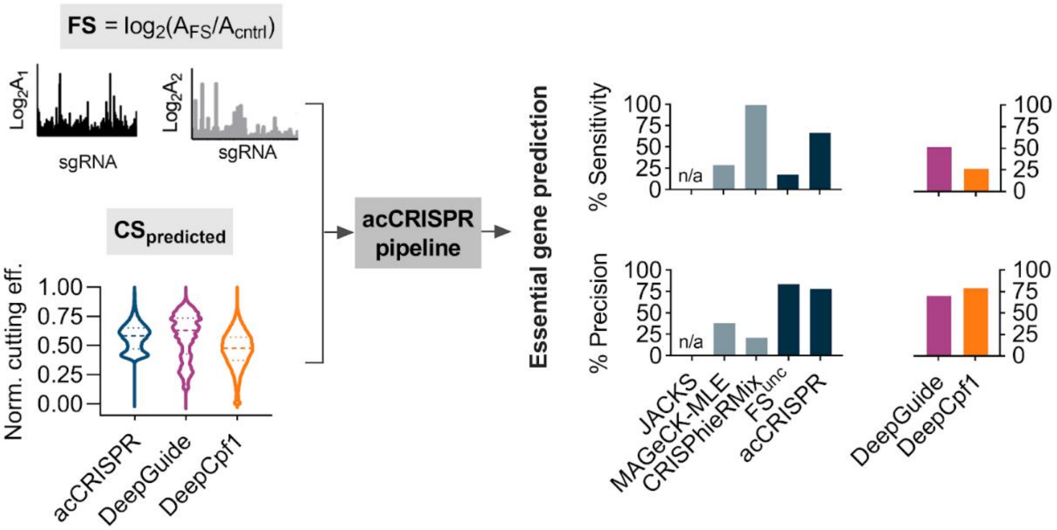


Fig. S2.4. Performance of acCRISPR on the Cas12a screening dataset with predicted sgRNA activities. Essential genes were determined with acCRISPR utilizing FS along with predicted sgRNA activities from DeepGuide [5] and DeepCpf1 [6]. The violin plot shows min-max normalized sgRNA activity distributions 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.

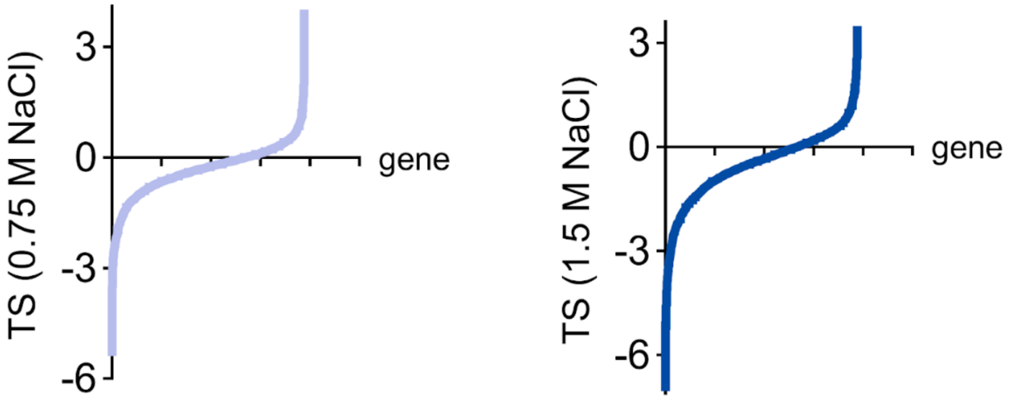


Fig. S2.5. acCRISPR corrected Tolerance Scores (TS) for salt tolerance screens. S-curves showing tolerance scores of genes at a CS threshold of 4.5 for the two salt stress conditions – 0.75 M NaCl (left) and 1.5 M NaCl (right).

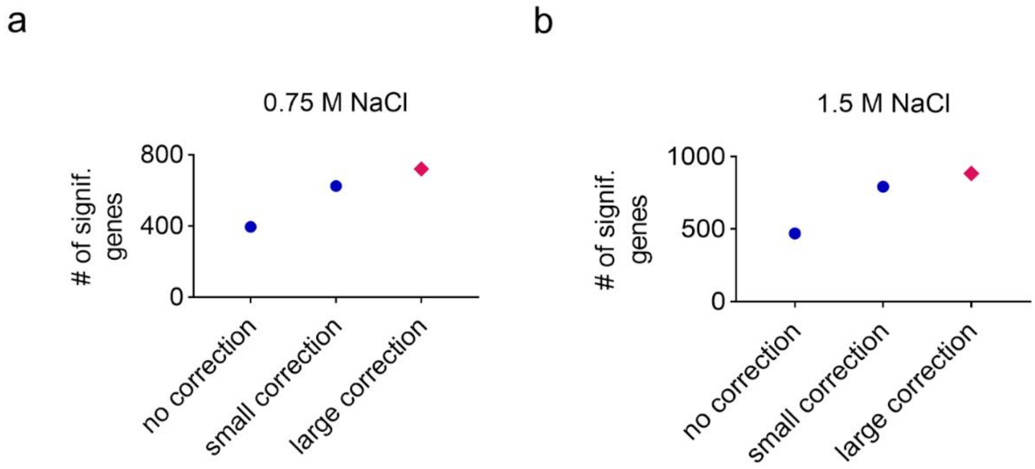


Fig. S2.6. Number of significant genes at different levels of activity correction for high salt tolerance screens. Dark blue points represent the number of significant genes predicted by acCRISPR without CS correction and with a small CS correction (CS-threshold = 1.0), while pink diamonds indicate the number of predicted significant genes with a large CS correction (CS-threshold = 4.5, i.e., optimum CS-threshold) for NaCl concentrations of (**a**) 0.75 M, and (**b**) 1.5 M.

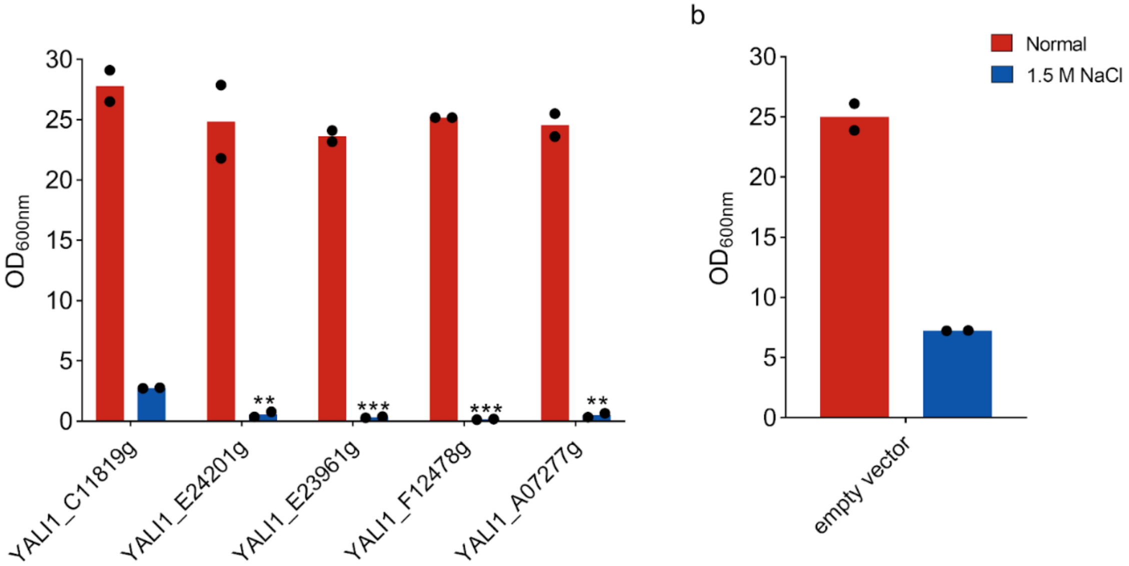


Fig. S2.7. Experimental validation of high salt tolerance genes from acCRISPR analysis. Bars represent final OD of (**a**) 5 single gene knockouts (*i.e.*, 4 significant genes and a non-significant gene, YALI1_C11819g), and (**b**) an empty vector control, grown in absence (normal; red bars) and presence (1.5 M NaCl; blue bars) of high salt conditions. Cells were grown in SD-Leu for 16 hrs post sgRNA transformation, followed by subculturing in fresh media (containing 1.5 M NaCl for the high salt condition) and growth for 4 days before measuring final OD. Bars indicate mean of two biological replicates (n = 2), and data points represent OD of each individual replicate in the respective sample (**p < 0.01, ***p < 0.001 ; one-tailed unpaired t-test).

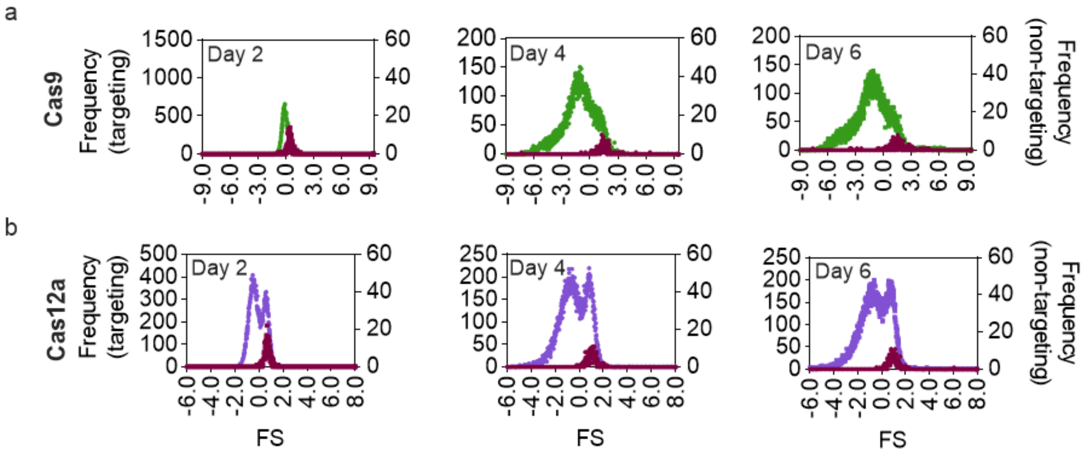


Fig. S2.8. CRISPR-Cas9 and -Cas12a FS distributions on days 2, 4 and 6. Green and purple distributions plotted on the left y-axis show FS of all targeting sgRNA in the library, while the dark red distributions plotted on the right y-axis represents the non-targeting populations. (**a**) Histogram of sgRNA FS values in the Cas9 dataset. (**b**) Histogram of sgRNA FS values in Cas12a dataset.

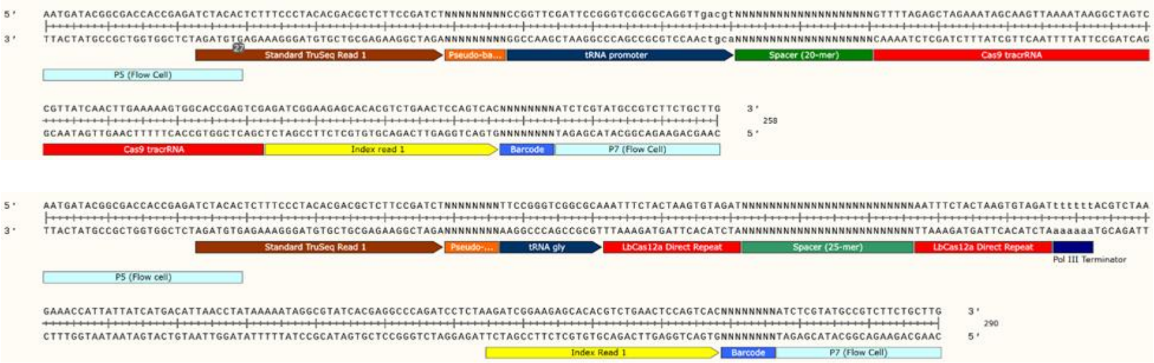

5'
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNNCCGGTTCGATTCCGGGTCGGCGCAGGTTgacgtNNNNNNNNNNNNNNNNNNNNGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTC
3'
TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGANNNNNNNNNGGCCAAGCTAAGGCCCAGCCGCGTCCAActgcaNNNNNNNNNNNNNNNNNNNNCAAAATCTCGATCTTTATCGTTCAATTTTATTCCGATCAG
27
Standard TruSeq Read 1
Pseudo-ba...
tRNA promoter
Spacer (20-mer)
Cas9 tracrRNA
P5 (Flow Cell)
CGTTATCAACTTGAAAAAGTGGCACCGAGTCGAGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG
3'
258
GCAATAGTTGAACTTTTTCACCGTGGCTCAGCTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNNNTAGAGCATACGGCAGAAGACGAAC
5'
Cas9 tracrRNA
Index read 1
Barcode
P7 (Flow Cell)
5'
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNNNNNTTCCGGGTCGGCGCAAATTTCTACTAAGTGTAGATNNNNNNNNNNNNNNNNNNNNNNNNNAATTTCTACTAAGTGTAGATttttttACGTCTAA
3'
TTACTATGCCGCTGGTGGCTCTAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGANNNNNNNNAAGGCCCAGCCGCGTTTAAAGATGATTCACATCTANNNNNNNNNNNNNNNNNNNNNNNNNTTAAAGATGATTCACATCTAaaaaaaTGCAGATT
Standard TruSeq Read 1
Pseudo-...
tRNA gly
LbCas12a Direct Repeat
Spacer (25-mer)
LbCas12a Direct Repeat
Pol III Terminator
P5 (Flow cell)
GAAACCATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCAGATCCTCTAAGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNNNATCTCGTATGCCGTCTTCTGCTTG
3'
290
CTTTGGTAATAATAGTACTGTAATTGGATATTTTTATCCGCATAGTGCTCCGGGTCTAGGAGATTCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTGNNNNNNNNTAGAGCATACGGCAGAAGACGAAC
5'
Index Read 1
Barcode
P7 (Flow Cell)

Fig. S2.9. 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.

Supplementary Tables

Table S2.1. 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 S2.2. 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 S2.3. Plasmids used for genome wide CRISPR screens.

Plasmid name	Reference	Function
pCpf1_yl	[7]	Plasmid for CRISPR-LbCas12a based gene editing in *Y. lipolytica*
pCRISPRyl (Addgene #70007)	[8]	Plasmid for CRISPR-Cas9 based gene editing in *Y. lipolytica*
pLbCas12ayl	This study and [5]	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)	[9]	Plasmid containing homology arms for integration of hrGFP into the A08 locus
pHR_A08_LbCas12a	This study and [5]	Plasmid containing homology arms for integration of LbCas12a into the A08 locus
pHR_A08_Cas9	[10]	Plasmid containing homology arms for integration of Cas9 into the A08 locus
pLbCas12ayl-GW	This study and [5]	Vector containing sgRNA expression cassette for cloning Cas12a sgRNA library. (Does not contain Cas12a expression cassette)
pCas9yl-GW	[10]	Vector containing sgRNA expression cassette for cloning Cas9 sgRNA library. (Does not contain Cas9 expression cassette)
pCRISPRyl_KU70	This study and [11]	CRISPR plasmid for the disruption of KU70

Table S2.4. 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
Sg-Seq	CTTCGACTCTAGAGGATCTGG

Table S2.5. 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 S2.6. 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	CAAGCAGAAGACGGCATACGAGATACCTAGCTGTGACTGGAGTTCAGACGTGTGCCTTCCG	AGCTAGGT

	ATCTTAGAGGATCTGGGCCTCGTGATAC	
ILU11-R	CAAGCAGAAGACGGCATACGAGATAGAGATGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	TCATCTCT
ILU12-R	CAAGCAGAAGACGGCATACGAGATCTGGACTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTAGAGGATCTGGGCCTCGTGATAC	AAGTCCAG

Table S2.7. Primers used for NGS fragment amplification (Cas9).

Primer name	Primer Sequence	Illumina Barcode (Reverse primer) / Pseudo-Barcode (Forward primer) for demultiplexing
Cr_1665	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGTCCGGTTCGATTCCGGGTC	^AGTCCG
Cr_1666	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTGTAGTCCGGTTCGATTCCGGGTC	^GTAGTC
Cr_1667	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGTAGTCCGGTTCGATTCCGGGTC	^CAGTAG
Cr_1668	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCCAGTAGTCCGGTTCGATTCCGGGTC	^TCCAGT
Cr_1669	CAAGCAGAAGACGGCATACGAGATTCGCCTTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTCAAG	CAAGGCGA
Cr_1670	CAAGCAGAAGACGGCATACGAGATATAGCGTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTCAAG	GACGCTAT
Cr_1671	CAAGCAGAAGACGGCATACGAGATGAAGAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTCAAG	ACTTCTTC
Cr_1672	CAAGCAGAAGACGGCATACGAGATATTCTAGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTCAAG	CCTAGAAT
Cr_1673	CAAGCAGAAGACGGCATACGAGATCGTTACCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTCAAG	TGGTAACG
Cr_1709	CAAGCAGAAGACGGCATACGAGATGTCTGATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTCAAG	CATCAGAC
Cr_1710	CAAGCAGAAGACGGCATACGAGATTTACGCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTCAAG	GTGCGTAA
Cr_1711	CAAGCAGAAGACGGCATACGAGATTTGAATAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCGACTCGGTGCCACTTTTTCAAG	CTATTCAA

Table S2.8. Parameters for bioinformatics tools on Galaxy [12] used in the analysis of NGS reads (Cas12a).

Tool	Version	Parameters*
FastQC	v0.11.8	Default settings
Cutadapt	Galaxy Version 1.16.6 [13]	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 S2.9. Parameters for bioinformatics tools on Galaxy [12] used in the analysis of NGS reads (Cas9).

Tool	Version	Parameters*
FastQC	v0.11.8	Default settings
Cutadapt	Galaxy Version 1.16.6 [13]	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.

Table S2.10. List of sgRNA (& associated cutting scores) used for validation of essential genes, non-essential genes, and significant genes for salt tolerance.

Gene	sgRNA ID (from Cas9 lib.)	sgRNA sequence	Cutting score
YALI1_B20188g	YALI1_B20188g_3	TTGCATCCTGATCGAAACCA	6.88
YALI1_D06665g	YALI1_D06665g_6	GGATGCTGCTACTTCCAAAT	5.73
YALI1_E17613g	YALI1_E17613g_6	CTTTGCACACCCCGTCAATT	7.14
YALI1_F08292g	YALI1_F08292g_3	GAACTCGTCAGCGAGCACGG	6.07
YALI1_F27686g	YALI1_F27686g_6	GCAGAAGAACCGCCTCACCA	6.90
YALI1_E23184g	YALI1_E23184g_3	CGAGTCGCCGACAACTGTAA	7.52
YALI1_A21345g	YALI1_A21345g_1	TCAATAGTAGCCTCAGACAA	6.80
YALI1_A03069g	YALI1_A03069g_5	TGCATCGGCGATATGTTCCA	6.10
YALI1_B00908g	YALI1_B00908g_5	GTTCTACGAGACCGATCACC	7.27
YALI1_C08600g	YALI1_C08600g_3	AATGGGGTCGAACGAAACGC	6.07
YALI1_D03952g	YALI1_D03952g_6	CTCCTGAGCGGCCTTCCACG	6.59
YALI1_D14276g	YALI1_D14276g_2	CTGGATCTCCAGCTGTACCG	6.35
YALI1_B03043g	YALI1_B03043g_2	AATGTCGCTCTGGTGAGTGA	6.64
YALI1_E18269g	YALI1_E18269g_3	ACACGCACTCAGTAAGGCAG	6.25
YALI1_F34105g	YALI1_F34105g_6	GAACGCCGTGATCATCGGAC	6.72
YALI1_E24201g	YALI1_E24201g_4	GACGTGGGCAAGAAAAAGGA	6.12
YALI1_C11819g	YALI1_C11819g_5	GTTTTGCCAGTTCCCCAACG	5.58
YALI1_A07277g	YALI1_A07277g_4	TGGCGGAGATCTAGATGTCG	7.11
YALI1_F10122g	YALI1_F10122g_1	CAGAAGGGAAAGTAGTACCG	5.90
YALI1_F09056g	YALI1_F09056g_5	CCGAGAAAACGGCCAAAGGG	5.86
YALI1_E23961g	YALI1_E23961g_2	AGGCTACTCGGGAGGAAACA	5.73
YALI1_F12478g	YALI1_F12478g_6	ACTTGTGCGCGCCTCCCACT	5.90

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

5. 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).

6. 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).

7. 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).

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.* **5**, 356–359 (2016).

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.* **6**, 402–409 (2017).

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

11. 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).

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

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

Chapter 3: Balanced training datasets improve deep learning-based prediction of CRISPR sgRNA activity

3.1 Abstract

CRISPR-Cas systems have transformed the field of synthetic biology by providing a versatile method for genome editing. The cleavage efficiency of CRISPR systems depends largely on the sequence of the constituent sgRNA, necessitating the development of computational methods for designing active sgRNAs. While deep learning-based models have shown promise in predicting sgRNA activity, the accuracy of prediction is primarily governed by the dataset used in model training. Here, we trained a convolutional neural network model and a large language model (LLM) on balanced and imbalanced datasets generated from CRISPR-Cas12a screening data for the yeast *Yarrowia lipolytica*, and evaluated their ability to predict high- and low-activity sgRNAs. We further tested whether prediction performance can be improved by training on imbalanced datasets augmented with synthetic sgRNAs. Lastly, we demonstrated that adding synthetic sgRNAs to inherently imbalanced CRISPR-Cas9 datasets from *Y. lipolytica* and *Komagataella phaffii* leads to improved performance in predicting sgRNA activity, thus underscoring the importance of employing balanced training sets for accurate sgRNAs activity prediction.

3.2 Introduction

CRISPR systems have emerged as a potent tool in enabling targeted genome editing in assays ranging from individual genetic perturbation experiments to high-throughput functional genetic screens [1–4]. Precise genomic edits created using CRISPR systems have facilitated identification of the genetic basis of phenotypes and discovery of previously unknown biological mechanisms in various model and non-model organisms for industrial, agricultural and therapeutic applications [5–10]. CRISPR systems achieve efficient targeted editing by utilizing two components – a Cas endonuclease that creates a double stranded break, and a single guide RNA (sgRNA) that guides the Cas enzyme to the targeted genomic locus [11,12]. Genome editing efficacy depends on several factors such as the sequence and nucleotide composition of the sgRNA, propensity of the sgRNA to form secondary structure, genomic context, and epigenetic features like chromatin accessibility and DNA methylation [13–16]. As a result, CRISPR systems often have a broad spectrum of activity, with only a limited fraction of sgRNA successfully generating a desired genetic manipulation, thus emphasizing the need for computational approaches to design sgRNAs.

A host of computational tools for CRISPR sgRNA design have been developed that possess the ability to predict sgRNA activity in prokaryotic and eukaryotic organisms using machine learning and deep learning approaches [17–20]. These methods use large datasets that link sgRNA sequence with Cas activity as training sets to capture generalizable patterns and features of sgRNAs, and in doing so generate design rules for maximizing sgRNA activity [21,22]. The composition of the training datasets used as input to these methods thus plays a critical role in determining the accuracy of activity predictions. Training datasets consisting of a large number of sgRNAs with a wide distribution of activity lead to more accurate predictions of activity compared to skewed datasets [18].

In this work, we evaluated the effect of training set composition on the performance of deep learning methods for sgRNA activity prediction. We began by training a deep convolutional neural network model, DeepGuide [18], on previously reported CRISPR-Cas12a data from *Y. lipolytica* [6,18]. The performance of a series of DeepGuide models trained on the original dataset and datasets skewed toward high- and low-activity sgRNAs was evaluated for prediction accuracy. We also implemented an LLM architecture, HyenaDNA [23], on the original and skewed datasets, and observed similar prediction performance. Finally, we investigated the effect of augmenting imbalanced training datasets with synthetic sgRNAs on the ability to recover predictive power lost during training on imbalanced datasets.

3.3 Results and Discussion

Balanced training sets enable accurate predictions of sgRNA activity

The CRISPR-Cas12a sgRNA dataset in *Y. lipolytica*, previously reported in ref [6,18], is an 8-fold coverage library containing 57,018 sgRNA targeting ~98% of the protein-coding genes in the PO1f strain. Unbiased design and screening of this library produced a dataset containing a well-balanced representation of high- and low-activity sgRNAs. The activity of each guide was determined using an experimental cutting score

(CS), computed as the $\log_2$ ratio of sgRNA abundance in a strain deficient in non-homologous end joining (NHEJ) to that in a Cas12a-expressing strain deficient in NHEJ [18,24]. Using this CS dataset, we trained DeepGuide[18] to predict CRISPR-Cas12a activity based on sgRNA sequence. The data was split into training and test sets in the ratio 90:10, with the training set consisting of 50,731 sgRNAs, comprising ~41% high-activity sgRNA and ~59% low-activity sgRNA (**Fig. S3.1**). Model training resulted in a mean Pearson's r of 0.596 (experimental CS vs DeepGuide-predicted CS), thus establishing a baseline of model performance for sgRNA activity predictions.

Imbalanced training sets result in poor prediction of sgRNA activity

In this experiment, we sought to quantify the impact of a reduced training set on model performance by randomly removing high- or low-activity sgRNAs from the original training set and evaluating the performance of DeepGuide when trained on the imbalanced datasets. While Pearson's r computed for the test set captures the overall accuracy of the model in predicting the CS of each sgRNA, it does not gauge the model's ability to correctly predict high- and low-activity sgRNAs as measured in experimental assays [25–27]. For this reason, we evaluated the performance of DeepGuide to accurately classify a set of experimentally-validated high- and low-activity sgRNAs identified from individual phenotype screening experiments [18]. Based on the DeepGuide predicted CS values along with the experimental high-/low-activity classification for each sgRNA, we computed the fraction of high-activity and low-activity sgRNAs correctly predicted by the model in terms of True Positive Rate (TPR) and 1 - False Positive Rate (1-FPR) respectively (see Methods for details on TPR, 1-FPR calculation from predicted CS).

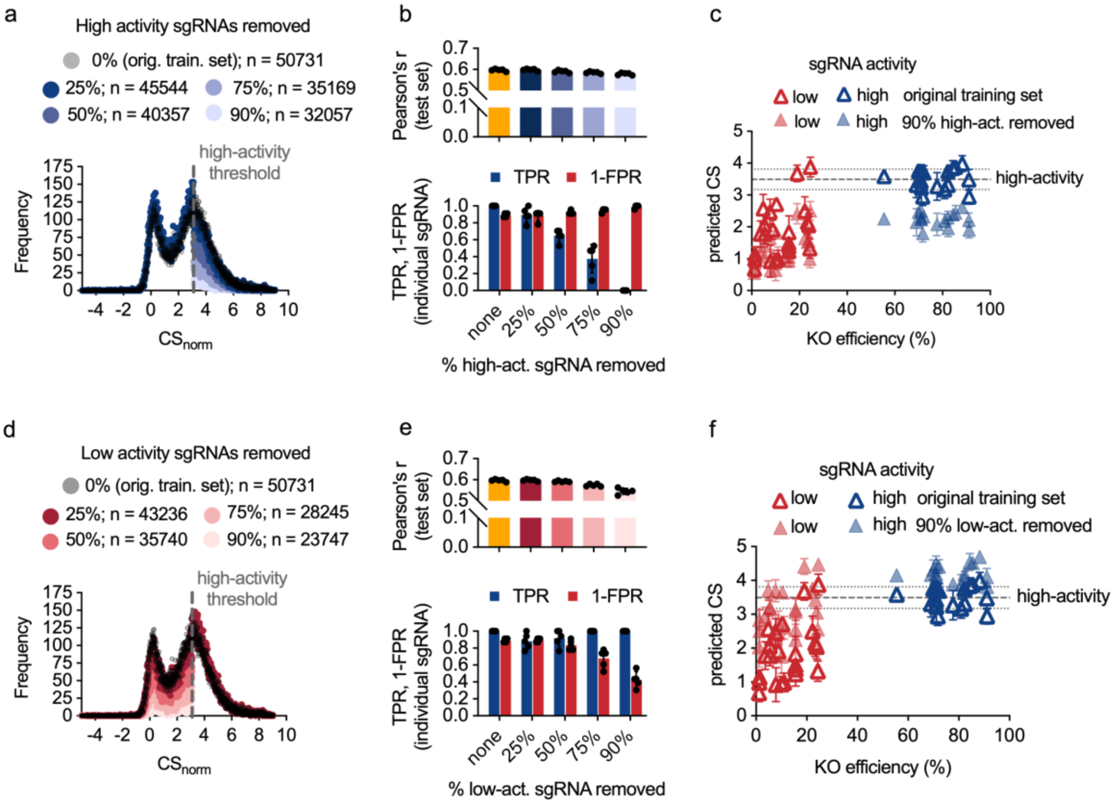


Fig. 3.1. DeepGuide performance with imbalanced CRISPR-Cas12a training datasets. (**a**), (**d**) Normalized cutting score (CS_{norm}) distributions of the Cas12a training dataset imbalanced by removing 0, 25, 50, 75, and 90% high- and low-activity sgRNAs along with the total number of sgRNAs (n) in every training set. (**b**), (**e**) Performance of DeepGuide models on the sgRNA test sets (Pearson's r), and high- and low-activity Cas12a sgRNAs from individual phenotype screening experiments (TPR, 1-FPR). Bars represent mean values of Pearson's r, TPR, and 1-FPR across five independent runs (n=5). Error bars indicate one standard deviation, and data points represent values from each individual run. (**c**), (**f**) Mean predicted CS of high-activity (KO efficiency >= 50%) and low-activity (KO efficiency < 50%) sgRNAs from individual phenotype screening experiments in *Y. lipolytica* when DeepGuide was trained on imbalanced datasets with 90% high-activity and 90% low-activity sgRNA removed. Data points represent mean values of predicted CS for experimental sgRNA with a given KO efficiency across five independent runs (n=5), and error bars indicate one standard deviation. Dashed line represents average predicted CS threshold for high-activity and dotted lines represent one standard deviation of the high-activity threshold.

We first generated imbalanced training datasets biased towards low-activity sgRNAs by removing 25, 50, 75 and 90% high-activity sgRNAs from the original training set (**Fig. 3.1a**). Upon training DeepGuide on these datasets, the performance on the test set was found to decrease slightly as more high-activity sgRNAs were removed, with mean Pearson's r ranging from 0.597 when 25% high-activity sgRNA were removed to 0.579 when 90% of the high-activity guides were excluded (**Fig. 3.1b**). Given that the Pearson's r for a balanced training set of 30,000 sgRNA was found to be higher (r = 0.588; **Fig. S3.2**) than that when 90% high-activity sgRNA were removed (r = 0.579; training set size ~ 32,000), the drop in Pearson's r can be attributed to a decrease in the share of high-activity sgRNAs in the training set. TPR was found to decline sharply as the percentage of high-activity sgRNAs removed increased from 25% to 90% – shrinking from a mean value of 0.894 to 0 when 25% and 90% high-activity sgRNAs were removed respectively, indicating a reduced ability to predict high-activity sgRNAs when the dataset is skewed toward low-activity guides (**Fig. 3.1b**). The decrease in TPR was accompanied by marginal increases in 1-FPR (mean value ranging from 0.887 when 25% high-activity sgRNAs were removed to 0.9914 when 90% were removed), reflecting the increasing share of low-activity sgRNAs in the training set. As the training datasets become more biased towards low-activity sgRNAs, the DeepGuide-predicted CS of experimental sgRNAs shift to lower values leading to fewer sgRNAs being predicted as high-activity (**Fig. 3.1c** and **Fig. S3.3a,b**).

We similarly removed 25, 50, 75, and 90% low-activity sgRNA from the original set to result in training data biased towards high-activity sgRNAs (**Fig. 3.1d**). Similar to the results with datasets biased toward low-activity sgRNAs, DeepGuide's performance on the test set decreased as more low-activity sgRNAs were removed (**Fig. 3.1e**; mean Pearson's r = 0.598 and 0.545 when 25% and 90% low-activity sgRNAs were removed respectively). Furthermore, as fewer low-activity sgRNA were retained in the training dataset, DeepGuide gradually lost the ability to accurately predict low-activity experimental data, with 1-FPR diminishing from 0.896 when 25% low-activity sgRNA were removed, to 0.426 when 90% low-activity sgRNA were removed (**Fig. 3.1e**). This result was coupled with a slight increase in TPR values (0.882 and 1 when 25% and 90% low-activity sgRNA were removed respectively). **Fig. 3.1f** and **Fig. S3.3a,c** show that the predicted CS of experimental sgRNAs shift to higher values with respect to the predicted CS obtained by training DeepGuide on the original dataset, ultimately causing fewer sgRNA to be predicted as low-activity.

To test whether alternative deep learning frameworks exhibit similar behavior in predicting sgRNA activity when trained on imbalanced datasets, we evaluated the performance of HyenaDNA [23], a large language model (LLM), trained on various datasets. For this analysis, we used the balanced training dataset and the skewed datasets with 50% and 90% high- and low-activity sgRNAs removed. Similar to the DeepGuide results, Pearson's r slightly decreased as more high- and low-activity sgRNAs were removed (**Fig. S3.4a**). Likewise, the TPR and 1-FPR showed a decrease that is commensurate with the fraction of high- and low-activity sgRNAs preserved in the training dataset (**Fig. S3.4b**). These results substantiate the critical importance of the training set composition in influencing sgRNA activity prediction accuracy, independent of the prediction model.

Augmenting imbalanced training sets with synthetic sgRNAs helps recover activity prediction power

To examine whether artificially re-balancing training sets improves prediction performance, we augmented imbalanced CRISPR-Cas12a training sets with synthetic high- or low-activity sgRNA. It is well-established that CRISPR endonuclease activity is less sensitive to mismatches between the sgRNA and target DNA in the PAM-distal region of the sgRNA compared to the PAM-proximal or seed region [28–30]. For CRISPR-Cas12a, the first 14 bp of a sgRNA from the 5' end comprise the seed region [28,31]. We thus generated synthetic sgRNAs by randomly sampling guides from the minority class in a given imbalanced training set (for a training set biased towards low-activity sgRNA, the minority class constitutes all high-activity sgRNAs within the training set, and vice-versa) and created new guides with random one nucleotide substitution in the non-seed region (base positions 15-25 from the 5' end) of the selected guide. Since the CRISPR-Cas12a library was designed by ensuring the uniqueness of the 14 bp sgRNA seed region in the genome [6,18], the generated synthetic sgRNA would always target the same genomic locus as the original sgRNA it was created from.

To re-balance training sets biased towards low-activity sgRNA, we augmented the datasets consisting of 50% and 90% high-activity sgRNAs removed with 20,000 and 28,000 synthetic high-activity sgRNA, respectively (**Fig. 3.2a**). DeepGuide training on these re-balanced datasets resulted in a small decrease in performance on the test set compared to that for the corresponding imbalanced training sets, with the mean Pearson's r dropping from 0.591 to 0.539, and from 0.579 to 0.496 when datasets with 50% and 90% high-activity sgRNAs removed were balanced with synthetic guides (**Fig. 3.2b**). The TPR for experimental high-activity sgRNAs, however, exhibited an increase when synthetic high-activity sgRNAs were added to the training sets (an increase of 0.223 and 0.459 for datasets with 50% and 90% high-activity sgRNA removed, respectively). For the dataset with 50% high-activity sgRNA removed, the recovery in performance yields predictions that closely match those achieved using the original training set (mean TPR = 0.870 for the balanced dataset, and 1 for the original dataset). The 1-FPR value, meanwhile, showed a small decrease when synthetic high-activity sgRNA were appended to the training sets, while still remaining above 0.85 for all datasets. **Fig. 3.2c** shows that the addition of synthetic high-activity sgRNAs to imbalanced training sets causes the predicted CS of experimental sgRNAs to shift to higher values, illustrating the recovery in high-activity guide prediction accuracy.

We next re-balanced training datasets with 50% and 90% low-activity sgRNAs removed by augmenting them with 6,000 and 18,000 synthetic low-activity sgRNAs, respectively (**Fig. 3.2d**). This resulted in minimal change in Pearson's r for DeepGuide predictions on the test set (an uptick of 0.004 and 0.015 for datasets with 50% and 90% low-activity sgRNA removed respectively; **Fig. 3.2e**). More importantly, the addition of synthetic sgRNAs led to a resurgence in 1-FPR for both datasets (an increase of 0.026 and 0.357 for datasets with 50% and 90% low-activity sgRNAs removed respectively), signifying that DeepGuide regains the ability to predict low-activity sgRNAs. It is noteworthy here that model performance for the dataset supplemented with synthetic sgRNAs after removing 90% of the low-activity population (1-FPR = 0.783, TPR = 0.976) is inferior to that for the original dataset (1-FPR = 0.887, TPR = 1) only by a small margin. **Fig. 3.2f** illustrates the shift in predicted CS of experimental sgRNAs to lower values upon addition of synthetic low-activity sgRNA to imbalanced training datasets.

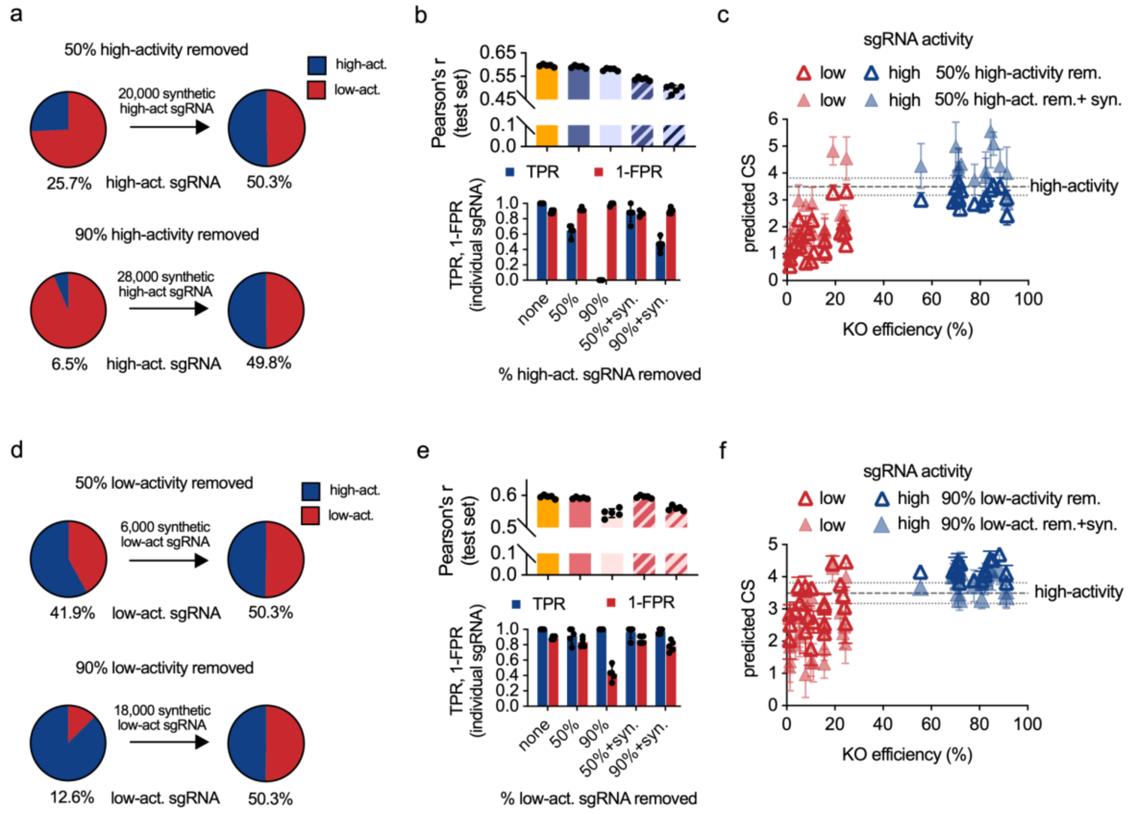


Fig. 3.2. DeepGuide performance with imbalanced CRISPR-Cas12a training datasets augmented with synthetic sgRNAs. (**a**), (**d**) Pie charts showing change in composition of imbalanced training sets skewed towards low- and high-activity sgRNAs after adding synthetic (**a**) high-activity, and (**d**) low-activity sgRNAs. (**b**), (**e**) Performance of DeepGuide models on the test set of sgRNAs (Pearson's r), and high-activity and low-activity Cas12a sgRNAs from individual phenotype screening experiments (TPR, 1-FPR), when trained using the original training set, imbalanced training sets obtained after removing 50% and 90% (**b**) high- and (**e**) low-activity sgRNAs, and re-balanced training sets obtained after adding synthetic (**b**) high- and (**e**) low-activity sgRNAs. Bars represent mean values of Pearson's r, TPR and 1-FPR across five independent runs (*n*=5), error bars indicate one standard deviation, and data points represent values from each individual run. (**c**), (**f**) Mean predicted CS of high-activity (KO efficiency >= 50%) and low-activity (KO efficiency < 50%) sgRNA from individual phenotype screening experiments in *Y. lipolytica* when DeepGuide was trained on balanced datasets containing synthetic (**c**) high- and (**f**) low-activity sgRNAs, with respect to the mean predicted CS of the same guides obtained upon training DeepGuide on the corresponding imbalanced datasets with 50% high-activity and 90% low-activity sgRNA removed. Data points represent mean values of predicted CS for experimental sgRNA with a given KO efficiency across five independent runs (*n*=5), and error bars indicate one standard deviation. Dashed line represents average predicted CS threshold for high-activity and dotted lines represent one standard deviation of the high-activity threshold.

We also explored variations of the current approach to generate synthetic sgRNA, and investigated their ability to improve prediction performance. The tested variations include: (i) penalizing the normalized CS of the sampled sgRNA by 4% (or (1/25)), assuming that the resulting synthetic sgRNA will have reduced activity due to a one nucleotide mismatch between sgRNA and target, (ii) sampling sgRNA by biasing towards sgRNA with extreme (high/low) normalized CS values, (iii) creating one nucleotide substitution in the sampled sgRNA by biasing towards terminal positions (*i.e.*, positions close to the 3' end), and (iv) creating two nucleotide substitutions in the non-seed region of the sampled sgRNA. Addition of synthetic sgRNA generated using the different methods resulted in a similar performance on the test set, and was not an improvement over the method shown in Figure 2 (**Fig. S3.5**). This similarity in performance was also observed for the experimental sgRNAs; the mean TPR values for training sets with synthetic high-activity sgRNAs range between 0.435-0.447 across all methods except method (ii), while mean 1-FPR values for training sets with synthetic low-activity sgRNA range between 0.722-0.852 across methods (**Fig. S3.5**). Overall, the similar performance of variant methods implies that the method used for generating synthetic sgRNA has no effect on the improvement in model performance.

Adding synthetic sgRNA to imbalanced CRISPR-Cas9 datasets improves low-activity sgRNA prediction

To assess the capability of the synthetic sgRNA-based approach in improving activity prediction on imbalanced training sets from other species and endonucleases, we implemented DeepGuide on CRISPR-Cas9 datasets from *Y .lipolytica* and *K. phaffii* previously reported in refs [32] and [33]. The *Y. lipolytica* Cas9 dataset is biased towards high-activity sgRNA; the set includes 67.3% high-activity sgRNA with a training set size of 19,953 (**Fig. 3.3a,b**). To alleviate this imbalance, we augmented the training set with

6,500 synthetic low-activity sgRNA by creating a 1 bp substitution in the non-seed region (base positions 1-8 from the 5’ end; see Methods). Since DeepGuide improves Cas9 activity predictions using nucleosome occupancy information [18], we provided occupancy scores for every sgRNA in addition to sgRNA sequence as input for training on the Cas9 datasets. When trained on the original and re-balanced training sets, DeepGuide was found to yield nearly similar values of Pearson’s r on the test set of sgRNA (**Fig. 3.3c**). Addition of synthetic sgRNA also resulted in an increase in 1-FPR from 0.053 to 0.493 for the original and re-balanced datasets respectively, but at the cost of a decrease in TPR from 1 to 0.656, **Fig. 3.3d**. **Fig. S3.6a** shows the predicted CS of experimental high- and low-activity sgRNA before and after adding synthetic sgRNA to the original training set.

The *K. phaffii* training set contains a disproportionately large number of high-activity sgRNA (73.7% high-activity sgRNA in a training set of 27,821 sgRNA, **Fig. 3.3e,f**), and was hence, re-balanced by adding 13,000 synthetic low-activity sgRNA. DeepGuide implementation on the training sets resulted in similar values of Pearson’s r (**Fig. 3.3g**). More prominently, when measuring performance on experimentally-validated high- and low-activity sgRNA from individual experiments [33], the addition of synthetic low-activity sgRNA led to an jump in 1-FPR from 0.042 to 0.232, accompanied by a small decrease in TPR from 1 to 0.815 (**Fig. 3.3h** and **Fig. S3.6b**).

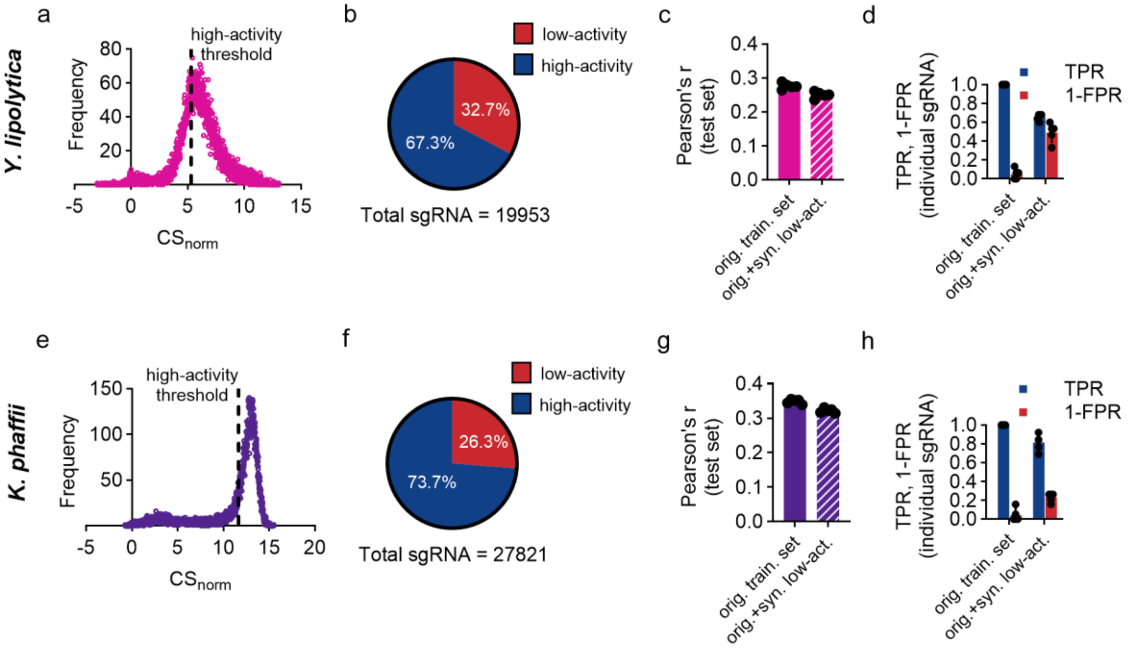


Fig. 3.3. Composition of the *Y. lipolytica* (top) and *K. phaffii* (bottom) CRISPR-Cas9 training sets and DeepGuide performance with the two datasets. (**a**), (**e**) Normalized cutting score (CS$_{norm}$) distributions of the original Cas9 training datasets for *Y. lipolytica* and *K. phaffii*. (**b**), (**f**) Pie charts showing the proportion of high- and low-activity sgRNAs in the original Cas9 training sets containing a total of 19,953 sgRNA for *Y. lipolytica*, and 27,821 sgRNA for *K. phaffii*. (**c**), (**g**) Performance of DeepGuide on the test set of sgRNA for *Y. lipolytica*, and *K. phaffii* when trained on the respective original training sets and re-balanced training sets obtained after adding synthetic low-activity sgRNA to the original sets. Bars represent mean Pearson's r across five independent runs (*n=5*), error bars indicate one standard deviation, and data points represent values from each individual run. (**d**), (**h**) DeepGuide performance on high- and low-activity Cas9 sgRNAs from individual experiments for *Y. lipolytica*, and *K. phaffii*. Bars represent mean values of TPR and 1-FPR across five independent runs (*n=5*), error bars indicate one standard deviation, and data points represent values from each individual run.

3.4 Conclusion

Deep learning models, while having shown to be effective in designing sgRNAs, depend significantly on the training set composition for accurate prediction of activity. Implementation of deep learning models on CRISPR-Cas datasets in this study shows that adding synthetic sgRNAs can improve performance with imbalanced datasets, but not to the level of balanced datasets. Ultimately, AI models result in best prediction performance when trained on datasets evenly representing both positive and negative biological outcomes, well-balanced datasets.

3.5 Methods

Processing Y. lipolytica and K. phaffii CRISPR-Cas12a and CRISPR-Cas9 sgRNA-CS data

Y. lipolytica sgRNA sequence and CS data for the CRISPR-Cas12a library was obtained from [18], while CRISPR-Cas9 datasets for *Y. lipolytica* and *K. phaffii* were obtained from [32] and [33] respectively. For all datasets, raw CS values of sgRNA were converted to normalized CS by subtracting the average CS of all non-targeting sgRNA in the respective libraries from the raw CS values of every sgRNA. For Cas12a data, the 25 bp sequences of sgRNA were extended to 32 bp sequences (25 bp spacer + 4 bp PAM + 1 bp context upstream of the PAM + 2 bp context downstream of the spacer) using custom Python scripts to map sgRNA to the *Y. lipolytica* CLIB89 genome (https://www.ncbi.nlm.nih.gov/assembly/GCA_001761485.1) [34] and obtain the upstream and downstream nucleotides. In case of the Cas9 datasets, the 20 bp sequences of sgRNA were extended to 28 bp sequences (20 bp spacer + 3 bp PAM + 2 bp context upstream of the spacer + 3 bp context downstream of the PAM) by mapping sgRNA to *Y. lipolytica* CLIB89 and *K. phaffii* GS115 (https://www.ncbi.nlm.nih.gov/assembly/GCA_000027005.1) [35] genomes.

For each dataset, the “sgRNA + PAM + upstream/downstream context” sequences and normalized CS data were then randomly split into training and test sets for the sgRNA activity prediction tools in the ratio 90:10. For *Y. lipolytica* CRISPR-Cas12a data, the original training set consisted of 50,731 sgRNA, while the test set comprised 5,637 sgRNA. Guides in the original training set were classified as high-activity and low-activity based on a high-activity threshold defined in [6], equivalent to a normalized CS of 3.10. The training and test sets for *Y. lipolytica* CRISPR-Cas9 data consisted of 19,953 and 2,217 sgRNA respectively, with a high-activity threshold equivalent to normalized CS of 5.30, as defined in [32]. Similarly, for *K. phaffii* Cas9 data, the training and test set sizes were 27,821 and 3,093 sgRNA respectively, with sgRNA having normalized CS greater than 11.66 deemed as high-activity sgRNA, based on the threshold defined in [33].

DeepGuide implementation

For *Y. lipolytica* datasets, DeepGuide (https://github.com/ucrbioinfo/deepguide_reborn) [18] was first pre-trained on the *Y. lipolytica* CLIB89 genome using a sequence length of 32 bp for Cas12a (guide_length: 32) and 28 bp (guide_length: 28) for Cas9 with 6 epochs (dg_one_pretrain_epochs: 6), followed by training on the *Y. lipolytica* Cas12a/Cas9 data with 10 epochs (dg_one_epochs: 10). For the *K. phaffii* Cas9 dataset, the pre-training was performed on the *K. phaffii* GS115 genome using 28 bp as the sequence length (guide_length: 28).

Both the pre-training and training steps were performed using a batch size of 64 (dg_one_pretrain_batch_size: 64, and dg_one_batch_size: 64) and a train:validation split of 70:30 (dg_one_pretrain_train_test_ratio: 0.7, and dg_one_train_test_ratio: 0.7). For Cas12a, the ‘cas’ parameter was set to ‘cas9_seq’, since only sgRNA sequence data was used for training. For Cas9 datasets, the value of ‘cas’ parameter was changed to ‘cas9_nucleosome’, since sgRNA nucleosome occupancy scores were used for training in addition to sequence data. Five independent runs were performed for each experiment.

HyenaDNA implementation

HyenaDNA (https://github.com/HazyResearch/hyena-dna) [23] was pre-trained on the *Y. lipolytica* CLIB89 genome using a sequence length of 32 bp (max_length: 32), train:val:test split of 80:10:10, model width of 32 (d_model: 32), depth of 2 layers (n_layer: 2), a learning rate of $6*10^{-4}$ (lr: 6e-4) and a global batch size of 1024 (global_batch_size: 1024) with 100 epochs (max_epochs: 100). Default values of all other parameters were used. Pre-training was carried out on 4 Nvidia A100 80GB GPUs (devices: 4).

For fine-tuning the model, the Cas12a training data was split into training and validation sets in the ratio 80:10 (train_len: 45094 for original training set), and a global batch size of 256 (global_batch_size: 256) was used. The model configuration, sequence length, and learning rate were kept unchanged from the pre-training step (d_model: 32, n_layer: 2, max_length: 32, lr: 6e-4). The fine-tuning step was also performed with 100 epochs (max_epochs: 100), using one Nvidia A100 80GB GPU (devices: 1), and the entire model was fine-tuned rather than freezing the weights of the pre-trained backbone (freeze_backbone: false). Default values of all other parameters were used. Five independent runs were performed for each experiment.

Generating synthetic sgRNA

Custom Python scripts were used to generate synthetic sgRNA by randomly sampling appropriate number of sgRNA from the pool of high-/low-activity sgRNA in the imbalanced training sets, and creating a 1 bp substitution for four of the five simulation methods, and 2 bp substitutions for one method, in the non-seed region (base positions 15-25 from the 5' end on the 25 bp spacer sequence for Cas12a sgRNA and positions 1-8 from the 5' end of the 20 bp spacer for Cas9 sgRNA [36,37]) of the sampled sgRNA.

In case of unbiased sampling with penalized CS for Cas12a sgRNA, the normalized CS of the synthetic guides was reduced by (1/25)th, or 4% compared to that of the original sgRNA to account for a possible reduction in sgRNA activity due to a 1 bp mismatch.

For biased sampling towards sgRNA with extremely high/low CS values, positive and negative exponential distributions were created for the range of normalized CS values for high-activity and low-activity guides respectively. For every simulated guide, a random value was sampled from this exponential distribution, and the normalized CS value closest to this sampled value and the corresponding sgRNA sequence were used to generate the synthetic guide.

For creating substitutions by biasing towards terminal positions on the sgRNA, the position for creating a substitution was sampled from an exponential distribution so that the probability of sampling terminal positions is higher compared to relatively central positions.

Computing nucleosome occupancy scores

Genome-wide nucleosome occupancy data for *Y. lipolytica* CLIB89 and *K. phaffii* GS115 genomes was obtained from MNase-seq datasets previously reported in [38] and [39] respectively. For every Cas9 sgRNA, an average occupancy score of the corresponding target locus was first computed by averaging the occupancy scores across all target bases, followed by normalizing the scores to values between 0 and 1 by dividing each average score by the highest average score in the respective dataset (*Y. lipolytica*/*K. phaffii*). The average normalized occupancy scores obtained for each sgRNA were then used to train DeepGuide alongside sgRNA sequence information.

Calculation of TPR and 1-FPR

Based on the predicted CS of sgRNA from individual phenotype screening experiments, every sgRNA was classified as high-activity or low-activity, which was different from the high-/low-activity classification based on experimental knockout efficiency. The predicted high-/low-activity classification was based on a p-value derived from a z-test of significance. Briefly, predicted CS values of experimental high-activity sgRNA (*i.e.*, sgRNA having knockout efficiency > 50%) obtained from the models trained on the original training sets were used to create a population of predicted CS of high-activity sgRNA for the respective datasets. Predicted CS values of experimental low-activity sgRNA obtained from the models trained on the original training sets, as well as predicted CS values of all (*i.e.*, experimental high-activity & low-activity) sgRNA in every subsequent activity prediction trial were compared to this population in a z-test of significance to determine if a given predicted CS value belongs to this population (p >

0.05; predicted high-activity sgRNA) or is significantly different from the population ($p < 0.05$; predicted low-activity sgRNA). The ability of a model to accurately predict sgRNA from individual experiments as high-activity and low-activity was measured using two metrics – True Positive Rate (TPR) and 1-False Positive Rate (FPR), respectively. TPR is defined as:

$$TPR = \left(\frac{No.of\ experimental\ high - activity\ sgRNA\ predicted\ to\ have\ high - activity}{Total\ no.of\ experimental\ high - activity\ sgRNA}\right)$$

Similarly, 1-FPR is calculated as:

$$1 - FPR = \left(\frac{No.of\ experimental\ low - activity\ sgRNA\ predicted\ to\ have\ low - activity}{Total\ no.of\ experimental\ low - activity\ sgRNA}\right)$$

Since the predicted CS values of experimental high-activity sgRNA from the model trained on the original set were used to generate the predicted high-activity population, all of these sgRNA were deemed to have high predicted activity, resulting in a TPR of 1 for the original training sets.

3.6 References

1. Przybyla, L. & Gilbert, L. A. A new era in functional genomics screens. *Nat. Rev. Genet.* **23**, 89–103 (2021).

2. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR–Cas9. *Nat. Rev. Genet.* **16**, 299–311 (2015).

3. Hart, T. *et al.* Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 Genes|Genomes|Genetics* **7**, 2719–2727 (2017).

4. Doench, J. G. Am I ready for CRISPR? A user's guide to genetic screens. *Nat. Rev. Genet.* **19**, 67–80 (2017).

5. Lupish, B. *et al.* Genome-wide CRISPR-Cas9 screen reveals a persistent null-hyphal phenotype that maintains high carotenoid production in Yarrowia lipolytica. *Biotechnol. Bioeng.* **119**, 3623–3631 (2022).

6. Ramesh, A. *et al.* acCRISPR: an activity-correction method for improving the accuracy of CRISPR screens. *Commun Biol* **6**, 617 (2023).

7. Jacobs, T. B., Zhang, N., Patel, D. & Martin, G. B. Generation of a Collection of Mutant Tomato Lines Using Pooled CRISPR Libraries. *Plant Physiol.* **174**, 2023–2037 (2017).

8. Liu, H.-J. *et al.* High-Throughput CRISPR/Cas9 Mutagenesis Streamlines Trait Gene Identification in Maize[OPEN]. *Plant Cell* **32**, 1397–1413 (2020).

9. Behan, F. M. *et al.* Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature* **568**, 511–516 (2019).

10. Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M. D. C. & Yusa, K. Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat. Biotechnol.* **32**, 267–273 (2013).

11. Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science* (2014) doi:10.1126/science.1258096.

12. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* (2012) doi:10.1126/science.1225829.

13. Javaid, N. & Choi, S. CRISPR/Cas System and Factors Affecting Its Precision and Efficiency. *Front. Cell Dev. Biol.* **9**, 761709 (2021).

14. Horlbeck, M. A. *et al.* Nucleosomes impede Cas9 access to DNA in vivo and in vitro. (2016) doi:10.7554/eLife.12677.

15. 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).

16. Přibylová, A., Fischer, L., Pyott, D. E., Bassett, A. & Molnar, A. DNA methylation can alter CRISPR/Cas9 editing frequency and DNA repair outcome in a target-specific manner. *New Phytol.* **235**, 2285–2299 (2022).

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

18. 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).

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

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

21. Sherkatghanad, Z., Abdar, M., Charlier, J. & Makarenkov, V. Using traditional machine learning and deep learning methods for on- and off-target prediction in CRISPR/Cas9: a review. *Brief. Bioinform.* **24**, bbad131 (2023).

22. Konstantakos, V., Nentidis, A., Krithara, A. & Paliouras, G. CRISPR–Cas9 gRNA efficiency prediction: an overview of predictive tools and the role of deep learning. *Nucleic Acids Res.* **50**, 3616–3637 (2022).

23. Nguyen, E. *et al.* HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *ArXiv* (2023).

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

25. Asuero, A. G., Sayago, A. & González, A. G. The Correlation Coefficient: An Overview. *Crit. Rev. Anal. Chem.* (2006) doi:10.1080/10408340500526766.

26. Berenson, G. S. Misleading correlations in clinical applications. *Clin. Chim. Acta* **40**, 266–268 (1972).

27. Westgard, J. O. & Hunt, M. R. Use and Interpretation of Common Statistical Tests in Method-Comparison Studies. *Clin. Chem.* **19**, 49–57 (1973).

28. Rabinowitz, R. & Offen, D. Single-Base Resolution: Increasing the Specificity of the CRISPR-Cas System in Gene Editing. *Mol. Ther.* **29**, 937–948 (2021).

29. Swarts, D. C., van der Oost, J. & Jinek, M. Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. *Mol. Cell* **66**, 221–233.e4 (2017).

30. Kim, H. *et al.* Enhancement of target specificity of CRISPR–Cas12a by using a chimeric DNA–RNA guide. *Nucleic Acids Res.* **48**, 8601–8616 (2020).

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

32. Robertson, N. R. *et al.* Optimized genome-wide CRISPR screening enables rapid engineering of growth-based phenotypes in Yarrowia lipolytica. *bioRxiv* 2024.06.20.599746 (2024) doi:10.1101/2024.06.20.599746.

33. Functional genomic screening in Komagataella phaffii enabled by high-activity CRISPR-Cas9 library. *Metab. Eng.* **85**, 73–83 (2024).

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

35. De Schutter, K. *et al.* Genome sequence of the recombinant protein production host Pichia pastoris. *Nat. Biotechnol.* **27**, 561–566 (2009).

36. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).

37. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233–239 (2013).

38. 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).

39. Liachko, I. *et al.* GC-Rich DNA Elements Enable Replication Origin Activity in the Methylotrophic Yeast Pichia pastoris. *PLoS Genet.* **10**, e1004169 (2014).

3.7 Supplementary Information

Supplementary Figures

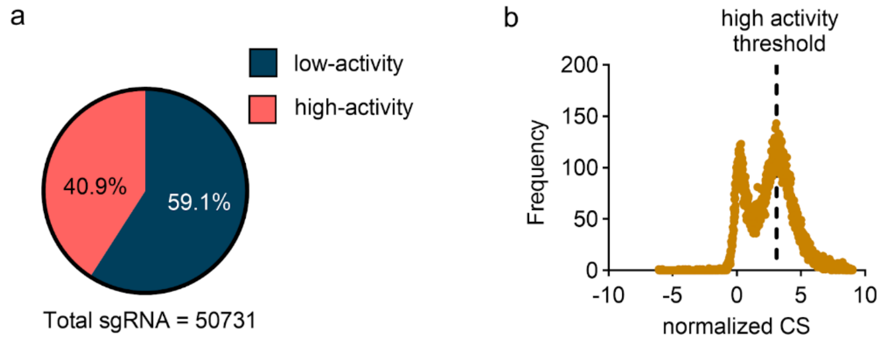


Fig. S3.1. Composition of *Yarrowia lipolytica* CRISPR-Cas12a training set. (**a**) Pie chart showing the proportion of high-activity and low-activity sgRNA in the *Y. lipolytica* CRISPR-Cas12a original training set containing a total of 50,731 sgRNA. (**b**) Distribution of normalized cutting scores (CS_{norm}) of sgRNA in the original Cas12a training set.

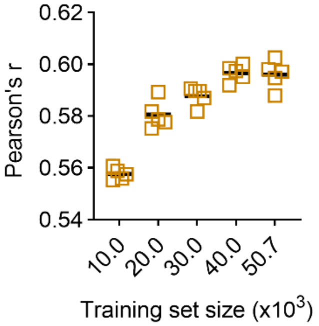


Fig. S3.2. DeepGuide performance as a function of CRISPR-Cas12a training set size. Performance of DeepGuide [1] on the Cas12a test set for different training set sizes. Horizontal bars represent mean values of Pearson’s r across five independent runs (*n=5*), and data points represent values from each individual run.

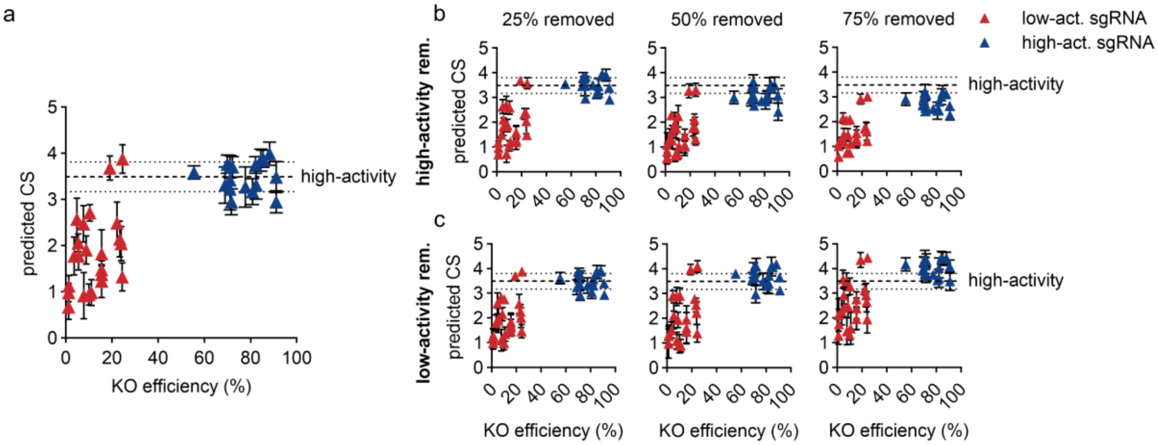


Fig. S3.3. DeepGuide CS predictions for individual high- and low-activity Cas12a sgRNA using different training sets. Mean predicted CS of high-activity (KO efficiency >= 50%) and low-activity (KO efficiency < 50%) sgRNA when DeepGuide was trained on (**a**) the original Cas12a training set, and (**b**), (**c**) imbalanced training sets with 25%, 50% and 75% (**b**) high- and (**c**) low-activity sgRNA removed. Data points represent mean values of predicted CS for experimental sgRNA with a given KO efficiency across five independent runs (*n=5*), and error bars indicate one standard deviation. Dashed line represents average predicted CS threshold for high-activity and dotted lines represent one standard deviation of the high-activity threshold.

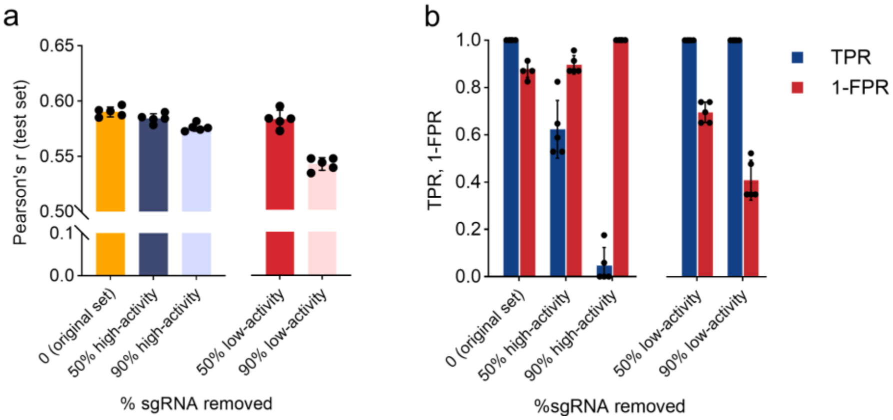


Fig. S3.4. Benchmarking performance of a large language model (LLM) on *Y. lipolytica* CRISPR-Cas12a dataset. Performance of HyenaDNA [2] on (**a**) the test set of sgRNA, and (**b**) high- and low-activity Cas12a sgRNA from individual phenotype screening experiments, when trained on the original Cas12a training set, and imbalanced training sets with 50% and 90% high- and low-activity sgRNA removed. Bars indicate mean values of Pearson's r, TPR and 1-FPR across five independent runs (*n*=*5*), error bars represent one standard deviation, and data points represent values from each individual run.

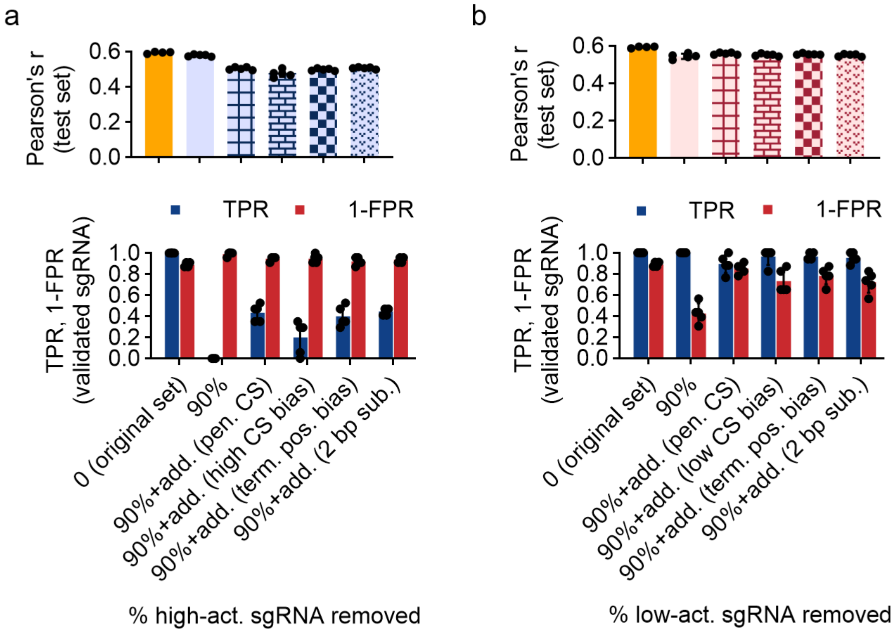


Fig. S3.5. DeepGuide performance with re-balanced training sets containing synthetic sgRNA generated using different variations of the unbiased method. Performance of DeepGuide models on the test set of sgRNA (Pearson's r), and high-activity and low-activity Cas12a sgRNA from individual phenotype screening experiments (TPR, 1-FPR), when trained using the original training set, imbalanced training sets obtained after removing 90% (**a**) high- and (**b**) low-activity sgRNA, and re-balanced training sets obtained after adding synthetic (**a**) high- and (**b**) low-activity sgRNA generated by:- penalizing normalized CS of synthetic sgRNA by 4% (pen. CS), sampling sgRNA by biasing towards sgRNA with extreme normalized CS values (high/low CS bias), creating a substitution in the sampled sgRNA by biasing towards terminal positions (term. pos. bias), and creating 2 bp substitutions in the non-seed region of the sampled sgRNA (2 bp sub.). Bars represent mean values of Pearson's r, TPR and 1-FPR across five independent runs (*n=5*), error bars indicate one standard deviation, and data points represent values from each individual run.

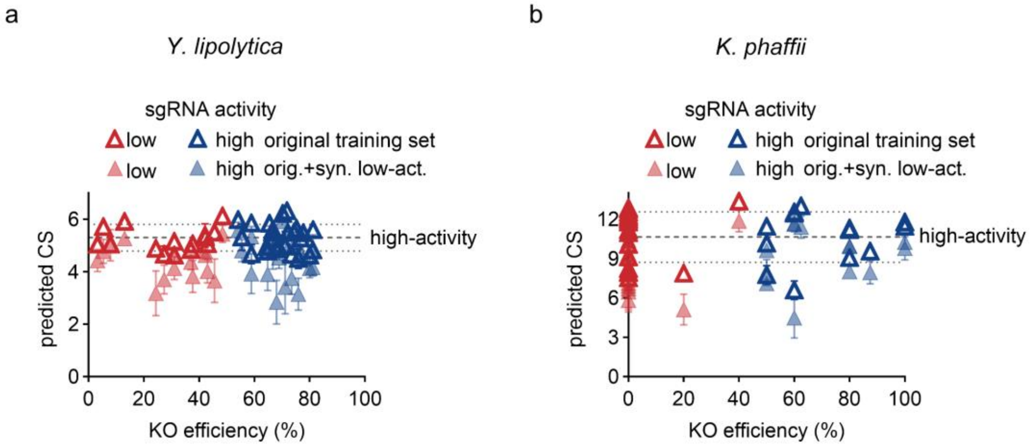


Fig. S3.6. DeepGuide CS predictions for individual high- and low-activity Cas9 sgRNA from *Y. lipolytica* and *K. phaffii*. Mean predicted CS of high-activity (KO efficiency >= 50%) and low-activity (KO efficiency < 50%) sgRNA when DeepGuide was trained on (**a**) *Y. lipolytica*, and (**b**) *K. phaffii* Cas9 datasets, before and after adding synthetic low-activity guides to the original set. Data points represent mean values of predicted CS for experimental sgRNA with a given KO efficiency across five independent runs (*n=5*), and error bars indicate one standard deviation. Dashed line represents average predicted CS threshold for high-activity and dotted lines represent one standard deviation of the high-activity threshold.

References

1. 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).

2. Nguyen, E. *et al.* HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. *ArXiv* (2023).

Chapter 4: Population genomics-guided engineering of phenazine biosynthesis in *Pseudomonas chlororaphis*

4.1 Abstract

The emergence of next-generation sequencing (NGS) technologies has made it possible to not only sequence entire genomes, but also identify metabolic engineering targets across the pangenome of a microbial population. This study leverages NGS data as well as existing molecular biology and bioinformatics tools to identify and validate genomic signatures for improving phenazine biosynthesis in *Pseudomonas chlororaphis*. We sequenced a diverse collection of 34 *Pseudomonas* isolates using short- and long-read sequencing techniques and assembled whole genomes using the NGS reads. In addition, we assayed three industrially relevant phenotypes (phenazine production, biofilm formation, and growth temperature) for these isolates in two different media conditions. We then provided the whole genomes and phenazine production data to a unitig-based microbial genome-wide association study (mGWAS) tool to identify novel genomic signatures responsible for phenazine production in *P. chlororaphis*. Post-processing of the mGWAS analysis results yielded 330 significant hits influencing the biosynthesis of one or more phenazine compounds. Based on a quantitative metric (called the phenotype score), we elucidated the most influential hits for phenazine production and experimentally validated them *in vivo* in the most optimal phenazine producing strain. Two genes significantly increased phenazine-1-carboxamide (PCN) production: a histidine transporter (ProY_1), and a putative carboxypeptidase (PS__04251). A putative

MarR-family transcriptional regulator decreased PCN titer when overexpressed in a high PCN producing isolate. Overall, this work seeks to demonstrate the utility of a population genomics approach as an effective strategy in enabling identification of targets for metabolic engineering of bioproduction hosts.

4.2 Introduction

The development of next-generation sequencing and CRISPR genome editing has enabled entire microbial genomes to be sequenced and manipulated, resulting in genome-wide metabolic engineering approaches often within non-traditional hosts. With further advancements in DNA sequencing technologies it is now economically feasible for a single research group to sequence small collections of tens to hundreds of microbial isolates, sometimes even in-house with portable sequencing devices. New metabolic engineering strategies could take advantage of this increasing accessibility of microbial whole-genome sequencing data and existing bioinformatics tools to analyze this data to identify metabolic engineering targets from a collection of genomes in a “pangenome”-wide or population genomics approach.

A first step in the design of a new metabolic engineering project is the selection of an appropriate host that natively exhibits a phenotype of interest. This work seeks to improve phenazine production in the bacterium *Pseudomonas chlororaphis* by identifying non-intuitive genetic targets from a collection of *P. chlororaphis* isolates as part of a population genomics approach to metabolic engineering. Phenazines are redox-active, often colorful, secondary metabolites with applications in agriculture as antifungal agents and potential applications as redox mediators in flow cell batteries and bioelectrochemical devices [1–3]. *P. chlororaphis* is a commercially available biocontrol species that would make a good potential phenazine production host as it natively produces multiple phenazine derivatives, is non-pathogenic to humans and plants, can utilize the inexpensive carbon source glycerol, and has available synthetic biology tools for genetic manipulation. Many strains also have traits that are detrimental to industrial bioprocessing, such as biofilm formation and low growth temperatures. Here, we sequenced the genomes of 34 *Pseudomonas* isolates, characterized their bioprocess-relevant phenotypes (phenazine production, biofilm formation, and growth temperature), and conducted microbial genome-wide association studies (mGWAS) to select an optimal host strain for phenazine production and identify genetic manipulations that increase phenazine biosynthesis.

P. chlororaphis has already been successfully engineered for phenazine production, with metabolic engineering works pursuing rational design strategies. Replacing genes within the phenazine biosynthesis operon can modulate final phenazine composition and allow non-native phenazines to be produced, including 1-hydroxyphenazine [4] and phenazine-1,6-dicarboxylic acid derivatives iodinin and 1,6-dimethoxyphenazine [5]). Regulation of the phenazine biosynthesis operon provides opportunities to improve phenazine production, including *phzR* and *phzI* which directly regulate expression through quorum sensing [6] and components of the Gac/Rsm pathway (*e.g. rpeA, rsmE, lon* protease, *psrA, parS, gacA* [4,7–9]) which indirectly interact with PhzR/PhzI in response to other environmental factors. Increasing carbon flux through the shikimate pathway, such as by overexpressing *aroB*, *aroD*, *aroE, ppsA*, and *tktA* [7,8], also improves phenazine production by increasing flux through phenazine biosynthesis. Collectively, these approaches combined with fermentation optimization have been able to produce grams per liter titers of phenazines, including 0.68 g/L of 2-hydroxyphenazine [9], 3.6 g/L of 1-hydroxyphenazine [4], and 11.45 g/L of PCN in *P. chlororaphis* [8].

Our population genomics approach uses mGWAS to identify metabolic engineering targets from our genomic and phenotypic data. GWAS correlate genomic and phenotypic datasets to identify causal genetic variants [10]. While GWAS are most commonly used to identify human disease risk factors, recent bioinformatics tools have been developed to adapt these studies to bacteria [11–13]. We input our sequenced genomes and phenotypic data into DBGWAS [13], a unitig-based mGWAS tool, to identify genomic loci that are significantly associated with phenazine production. This approach requires no prior knowledge of relevant biosynthetic pathways and could identify previously unknown targets for metabolic engineering throughout a single genome and the genomes of a population of isolates. We further sought to experimentally validate the top mGWAS hits by overexpressing associated genes and measuring phenazine production with respect to the wildtype control. Genes associated to hits identified in this study present new targets for strain engineering of *P. chlororaphis* to improve phenazine bioproduction.

4.3 Results and Discussion

Curating a P. chlororaphis strain collection

We began our population genomics approach to metabolic engineering by collecting a library of *P. chlororaphis* strains, a known overproducer of phenazines. We purchased all unique strains that were accessible to us from international culture collections, resulting in 26 strains purchased from three culture collections (**Table S4.1**). Eight of these strains had multiple colony morphologies present which appeared to vary in pigment production. Because these variants could differ in the phenotypes of interest and consequently may have associated genetic variation, each of these variants was treated as a separate isolate with the strain number indicating the original culture collection strain designation followed by a superscript ^{1} and ^{2} arbitrarily assigned to the differing morphologies. 16s rRNA sequencing identified 33 isolates as *P. chlororaphis* and one as *P. synxantha* for a total of 34 *Pseudomonas* isolates used in this study.

Phenotyping for phenazine production, biofilm formation, and growth temperature

For the phenazine biosynthesis phenotyping, we quantify the four phenazine compounds naturally produced by *P. chlororaphis*: 2-hydroxyphenazine (2-HP), 2-hydroxyphenazine-1-carboxylic acid (2-HPCA), phenazine-1-carboxylic acid (PCA) and phenazine-1-carboxamide (PCN) (**Fig. 4.1**) [14]. In phenazine-producing pseudomonads, the core phenazine biosynthesis operon is responsible for synthesizing PCA which serves as the precursor for other phenazine derivatives. *P. chlororaphis* strains typically produce either PCN or a combination of 2-HPCA and 2-HP depending on whether *phzH* or *phzO*, respectively, is present and functional.

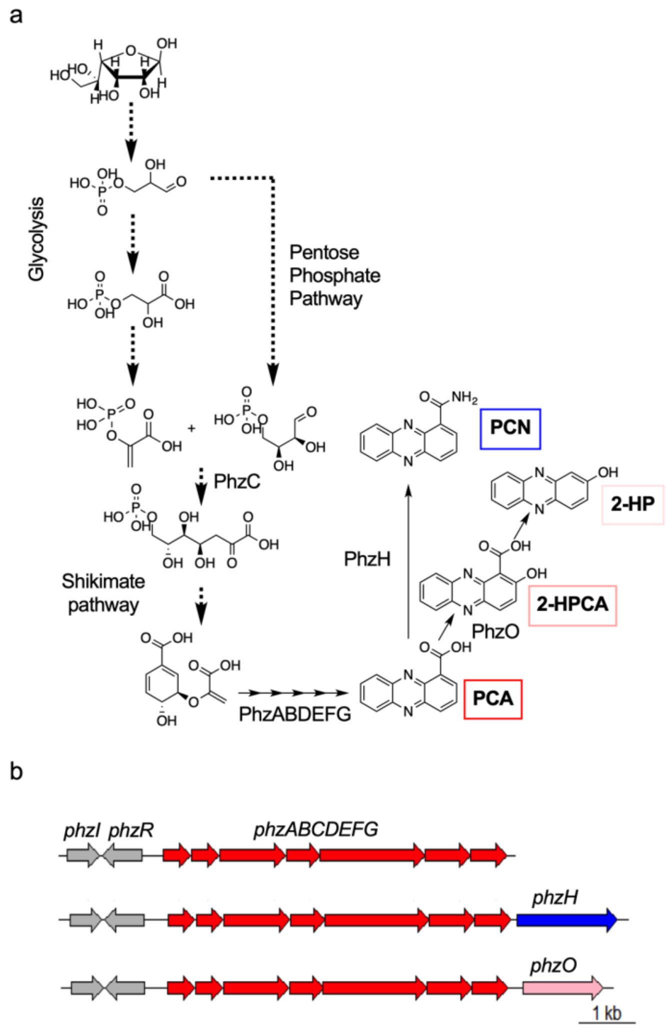


Fig. 4.1. Phenazine biosynthesis pathway and operon in *P. chlororaphis*. (**a**) Phenazine biosynthesis pathway. *P. chlororaphis* naturally produces 4 phenazines: 2-hydroxyphenazine (2-HP), 2-hydroxyphenazine-1-carboxylic acid (2-HPCA), phenazine-1-carboxylic acid (PCA) and phenazine-1-carboxamide (PCN). Chorismate from the shikimate pathway is converted into PCA. PCA can be converted into PCN or 2-HPCA by PhzH or PhzO, respectively. 2-HP is a byproduct of the spontaneous decarboxylation of 2-HPCA. (**b**) In fluorescent pseudomonads, phenazines are produced by a highly conserved core phenazine biosynthesis operon. *phzI/phzR* encodes a two-component quorum-sensing system which regulates the expression of the *phz* operon *phzABCDEFG*, which is responsible for the production of PCA. Some strains of *P. chlororaphis* contain *phzO*, the protein product of which converts PCA into 2-HPCA (which spontaneously decomposes into 2-HP), while others contain *phzH*, whose gene product converts PCA into PCN. *phzH* or *phzO* occur immediately downstream of *phzG* in the biosynthesis operon.

We first characterized phenazine production in King's Media B (KMB), the standard culture media for fluorescent pseudomonads (**Fig. 4.2**). Under these conditions, fewer than half the isolates produced more than 10 mg/L of phenazines. These low titers suggest that phenotyping in KMB may underestimate the phenazine production capacity of our strain collection. To improve phenazine production and consequently the quality of this dataset for our mGWAS analysis, we supplemented KMB with 100 µM ferric iron, which has been previously reported to enhance phenazine production in some strains of

P. chlororaphis [15,16]. KMB+Fe media improved phenazine production in 24 isolates, and 10 isolates did not produce significant phenazines in either medium. Due to its positive effects for most of the strains and its neutral effects on the remaining strains, we did additional phenotyping in KMB+Fe as well as KMB.

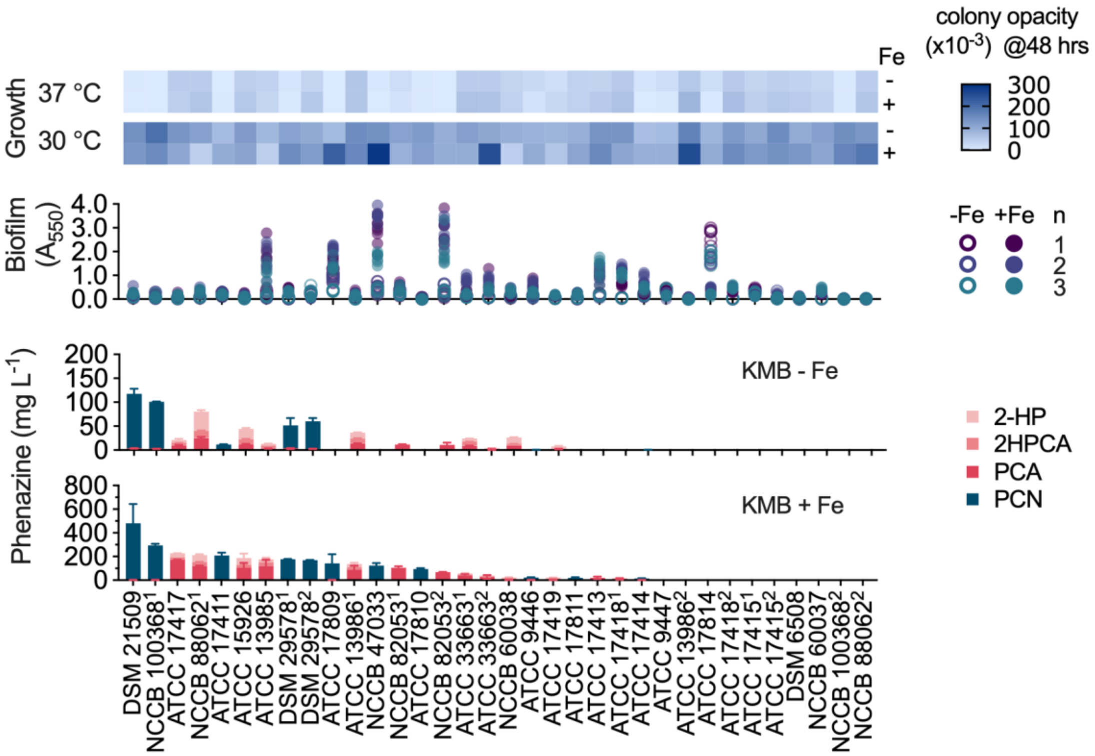


Fig. 4.2. Phenazine production, biofilm formation, and growth temperature phenotyping for all isolates used in this study. All phenotyping data was collected after 48 hours of culture in either King's Media B (-Fe) or King's Media B + 100 µM Fe^{3+} (+Fe). 2-hydroxyphenazine (2-HP), 2-hydroxyphenazine-1-carboxylic acid (2-HPCA), phenazine-1-carboxylic acid (PCA) and phenazine-1-carboxamide (PCN) were quantified using HPLC. The PCN-producers primarily produced PCN, with only very small amounts of the PCA precursor detected (< 5 mg/L). For phenazine production, bars indicate average of 3 replicates, and error bars represent one standard deviation. For growth temperature, the heat map shows a measure of colony growth on solid media (opacity (*10^{-3}); higher opacity indicates large and more dense colonies). For biofilm formation, each data point (A_{550}, which is indicative of biofilm formation) represents a separate biological replicate, which is the average of 8 technical replicates.

Out of all strains we characterized, strain DSM 21509 was found to produce significantly higher titers of PCN (477 ± 163 mg/L; one-way ANOVA followed by Tukey's test, $p < 0.05$) than all other strains in KMB + Fe. DSM 21509 was also one of the strains producing significantly higher PCN titers in KMB, along with NCCB 100368[1] (> 95 mg/L for both strains; $p < 0.001$). Strain DSM 21509 was thus deemed as the best host strain overall for PCN production. DSM 21509 is the type strain of *P. chlororaphis* subsp. *piscium*. This strain was isolated from the intestine of a European perch from Lake Neuchâtel, Switzerland, in 2005 [17]. Strains ATCC 17417, NCCB 88062[1], ATCC 15926 and ATCC 13985 were found to produce significantly higher titers of combined PCA/2-HPCA/2-HP in KMB+Fe (> 150 mg/L for each strain; $p < 0.05$) compared to the remaining strains, while strain NCCB 88062[1] alone produced a significantly higher titer of combined PCA/2-HPCA/2-HP in KMB (80 ± 6 mg/L; $p < 0.0001$). Strain NCCB 88062[1] was therefore considered the most optimum strain for production of PCA and its derivatives. This strain originated from the Netherlands and was deposited into NCCB in 1988 (https://wi.knaw.nl/page/NCCB_strains_display/24262). This phenazine production data for both media conditions was used as input for the mGWAS analysis.

In addition to phenazine production, we also characterized growth temperature and biofilm formation for all strains in both media conditions and used these phenotypic datasets to assess the potential of each isolate as a biotechnology host [18]. To identify strains which could grow at common bioprocessing temperatures, we characterized growth for all strains at 30 °C and 37 °C. These temperatures are relevant for *P. chlororaphis* since it is typically cultured at 28-30 °C, and some strains have been reported to grow at a maximum of 37 °C [19,20]. While some other fluorescent pseudomonads like *P. aeruginosa* can grow well and produce phenazines at 37 °C, there are no reports of phenazine production in *P. chlororaphis* at this temperature. In order to measure growth at 30 °C and 37 °C, strains were grown on solid media (KMB and KMB+Fe) at the two temperatures and the opacity of colonies (*i.e.*, the brightness of every colony pixel relative to its background, summed over the entire colony size) was measured after 48 hours of growth (see Materials and Methods for experimental details). At 37 °C, the average opacity of strains was relatively lower compared to that at 30 °C, indicating worse growth at 37 °C (**Fig. 4.2**; see **Fig. S4.1** for growth curves of all strains across 3 days in each condition). We also observed no colorful pigments on colonies at 37 °C, indicating little to no phenazine production, which agrees with the literature observations for *P. chlororaphis* [19]. The strains which can grow at 37 °C could be useful to pursue as hosts for other products but likely not for phenazines. Therefore 30 °C was selected as the fermentation temperature for this study, as all isolates could grow at the lower temperature on both media.

Biofilm formation is a phenotype which affects a host's ability to produce a desired product through altered cellular metabolism and growth kinetics and can be engineered within pseudomonads [21]. Given this, we characterized biofilm formation to determine whether future metabolic engineering efforts would be necessary to alter biofilm production in the desired host strain. For this study, we chose to minimize biofilm formation because high biofilm formation may be difficult to easily culture in a planktonic system and/or clean from industrial bioreactors. In KMB, only 2 strains produced noticeable biofilm. Biofilm formation did increase in KMB+Fe, but only 5 strains had enough biofilm for it to be visibly noticeable when handling the liquid culture. Strain DSM 21509 and PCN were selected as the production strain and desired phenazine product, respectively, because this strain produced the highest overall phenazine titers, which were 99.2% PCN in KMB+Fe. Additionally, DSM 21509 had low biofilm formation in both tested media conditions making it favorable to work with.

Genome sequencing, assembly, and annotation

We sequenced all strains with both Illumina and Oxford Nanopore technologies as each technology generates reads which vary in length and accuracy, therefore affecting the quality of the resulting assemblies. Using each read set separately or together (in a hybrid approach), we assembled genomes with different assembly algorithms (*i.e.*, SPAdes, Unicycler, Flye) to determine which algorithm and combinations of parameters yield the best assemblies. The summary statistics (*i.e.*, number of contigs, L_{50}, N_{50}, assembly length, GC content, number of CDS and BUSCO score) were compared to assess genome contiguity and accuracy and thus to select the optimal genome assemblies for the mGWAS analysis (**Fig. S4.2**).

Table 4.1. Summary statistics for final genome assemblies. For each isolate, the assembly length, number of contigs, N_{50}, L_{50}, GC% and BUSCO score are reported. These assemblies were used as input to mGWAS analysis. The number of CDS was tallied from the Prokka genome annotations, and the BUSCO score was calculated as the percentage of complete and single copy BUSCOs present in each genome using the BUSCO algorithm. All other statistics were generated from QUAST. These genome assemblies are available at NCBI with the listed accession numbers. All strains are *P. chlororaphis*, except ATCC 17413 (marked with *), which is *P. synxantha*.

Strain	Total length (bp)	Contigs	N_{50}	L_{50}	GC (%)	CDS	BUSCOs (%)	NCBI Accession Number
ATCC 13985	7 024 010	10	4 636 000	1	62.7	6251	99.1	JAQZQZ000000000
ATCC 13986[1]	6 675 284	2	6 636 555	1	63.0	5926	99.5	JAQZQY000000000
ATCC 13986[2]	6 682 756	2	6 644 045	1	63.0	5949	99.5	JAQZQX000000000
ATCC 15926	6 763 921	1	6 763 921	1	62.9	5998	99.4	CP118156
ATCC 17411	7 212 419	1	7 212 419	1	62.5	6366	99.2	CP118155
ATCC 17414	6 807 169	1	6 807 169	1	63.0	6048	99.5	CP118154
ATCC 17415[1]	6 664 157	1	6 664 157	1	63.0	5887	99.4	CP118147
ATCC 17415[2]	6 664 503	1	6 664 503	1	63.0	5884	99.2	CP118146
ATCC 17417	6 746 536	1	6 746 536	1	62.9	5954	99.1	CP118145
ATCC 17418[1]	6 883 267	1	6 883 267	1	62.8	6075	99.5	CP118144
ATCC 17418[2]	6 881 643	1	6 881 643	1	62.8	6074	99.5	CP118143
ATCC 17419	6 608 598	5	4 662 896	1	62.7	5919	99.0	JAQZQW000000000
ATCC 17809	7 020 903	1	7 020 903	1	62.4	6223	99.0	CP118142
ATCC 17810	6 863 056	2	6 791 445	1	62.7	6074	99.4	JAQZQV000000000
ATCC 17811	7 189 114	1	7 189 114	1	62.4	6422	99.4	CP118153
ATCC 17814	6 807 913	1	6 807 913	1	63.0	6050	99.4	CP118141
ATCC 33663[1]	7 109 352	1	7 109 352	1	62.9	6281	99.1	CP118152
ATCC 33663[2]	7 108 820	1	7 108 820	1	62.9	6284	99.2	CP118140
ATCC 9446	6 637 791	1	6 637 791	1	63.0	5909	99.2	CP118151
ATCC 9447	6 807 068	3	6 677 872	1	63.0	6048	99.4	JAQZQU000000000
DSM 21509	7 064 975	1	7 064 975	1	62.7	6246	99.1	CP118150
DSM 29578[1]	7 216 947	1	7 216 947	1	62.5	6378	99.1	CP118139
DSM 29578[2]	7 216 571	1	7 216 571	1	62.5	6380	99.1	CP118138
DSM 6508	7 915 166	3	7 476 725	1	62.5	7222	99.4	JAQZQT000000000
NCCB 100368[1]	6 870 522	2	6 455 838	1	62.8	6010	99.2	JAQZQS000000000
NCCB 100368[2]	6 870 415	2	6 455 628	1	62.8	6010	99.1	JAQZQR000000000
NCCB 47033	7 221 530	1	7 221 530	1	62.4	6378	99.2	CP118137
NCCB 60037	6 977 278	1	6 977 278	1	62.7	6209	99.1	CP118149
NCCB 60038	6 979 353	1	6 979 353	1	62.7	6214	99.1	CP118136
NCCB 82053[1]	6 763 242	2	6 274 577	1	62.9	6002	99.4	JAQZQQ000000000
NCCB 82053[2]	6 762 156	1	6 762 156	1	62.9	6002	99.4	CP118135
NCCB 88062[1]	7 025 460	2	6 660 177	1	62.8	6233	99.1	JAQZQP000000000
NCCB 88062[2]	6 923 225	1	6 923 225	1	62.8	6138	98.2	CP118148
*ATCC 17413	6 147 644	1	6 147 644	1	60.0	5459	99.9	CP118134

Contiguity statistics (number of contigs, N_{50}, L_{50}) describe the degree of fragmentation of an assembly. The number of contigs, or assembly fragments, should ideally approach one to accurately represent bacterial genomes with a singular circular chromosome, as is expected for *P. chlororaphis*. Genome completeness was assessed by implementing the BUSCO algorithm to calculate the percentage of expected complete and single copy orthologs which are present in each strain. The hybrid assemblies created with Unicycler were selected as the final assemblies due to their high contiguity and completeness metrics (L_{50} = 1 and BUSCO score >98% for all strains). Summary statistics for each of the final genomes, including total length, number of annotated CDS, BUSCOS scores, and assembly metrics are presented in **Table 4.1**. Notably, the vast majority of assemblies resulted in a single contig (22 *P. chlororaphis* and 1 *P. synxantha*), nine assembled into three or less contigs, one produced five contigs, and only one had ten contigs. Combined with the high BUSCO scores, the low number of contigs is indicative of high quality, complete genomes across our strain collection.

Assembling the P. chlororaphis pangenome

While the 33 *P. chlororaphis* isolates are members of the same species, their gene content varies among isolates. Comparing the assembly summary statistics (Table 1) reveals a wide range of assembly size (6.6 Mbp to 7.9 Mbp) and number of CDS (5884 to 7222 annotated CDS) so we decided to assemble the pangenome to gain further insight into these genomic differences. We input the Prokka annotations for the final *P. chlororaphis* genomes into the algorithm PEPPAN [22] to calculate the pangenome, the total gene content of the strains. The pangenome contains 11527 total CDS and 4406

CDS common to all strains (**Fig. 4.3a**). This translates to 61-75% of the CDS in each genome being common to all strains, the core *P. chlororaphis* genome. The remaining CDS are members of the accessory genome (genes present in some strains and absent in others), which corresponds to the majority of this pangenome (7121 genes). These accessory genes are found in a relatively small number of strains, while all strains contain the 4406 core genome (**Fig. 4.3b**).

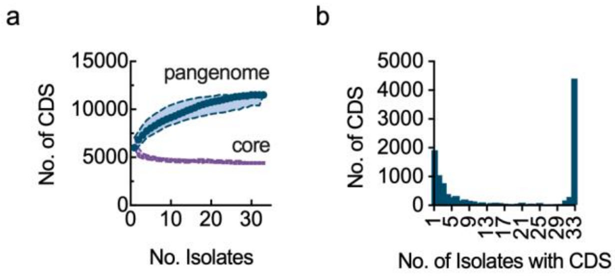


Fig. 4.3. Summary of the pangenome constructed from the final hybrid genome assemblies. (**a**) The pangenome rarefaction curve shows how the total pangenome size (blue) increases and the core genome size (purple) decreases as isolates are added to the pangenome. (**b**) Histogram showing how many CDS are only found in a specific number of strains. Core genes present in almost all strains or accessory genes only found in a few are present with the highest frequencies.

In addition to the genomic variation among the strains, we observed considerable phenotypic variation as shown in Fig. 2. Most notably, strains that produced PCN did not accumulate significant quantities of PCA or other phenazine derivatives. One of the key differences amongst these strains was the presence or absence of *phzH* and *phzO*; those that produced PCN contained *phzH*, while those that produced 2-HPCA contained *phzO*, as expected (**Fig. 4.4**). This split amongst the population is observable in a similarity tree based on accessory gene content, which clustered the isolates containing *phzH* separately from those that contain *phzO*. In the pre-genomic era the phenotypic differences were the basis of classification; members of these two groups would likely have been classified as different *Pseudomonas* species (*e.g.*, green pigment-producers as *P. chlororaphis*, yellow-orange pigment-producers as *P. aureofaciens* or *P. aurantiaca*; [23]). Classification in this way would incorrectly separate the groups, as our 16s and genomic sequencing shows that all 33 strains are *P. chlororaphis* with genetic variation driving the naturally produced phenazines. Collectively, the pangenome represents a large number of potential metabolic engineering targets that will be analyzed in our mGWAS.

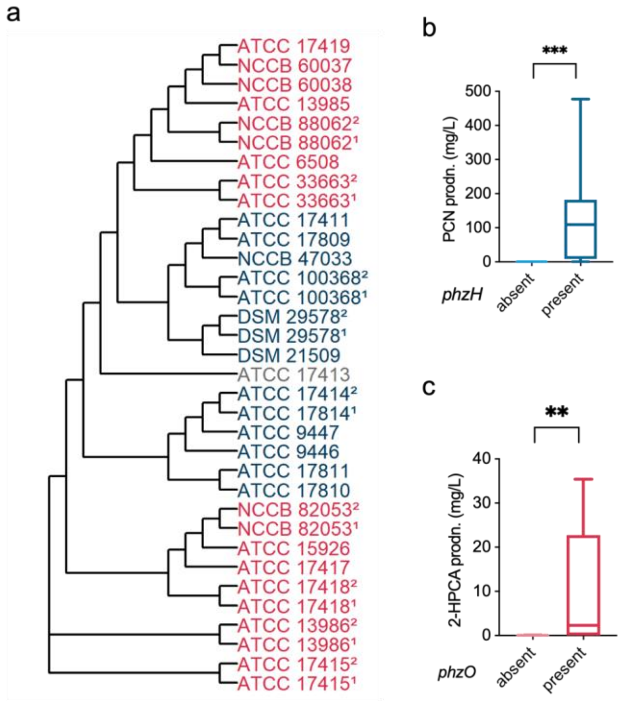


Fig. 4.4. Categorizing isolates based on presence and absence of genes *phzH* and *phzO*. (**a**) Tree based on accessory gene content among the *P. chlororaphis* pangenome. Strains containing *phzO* (red) and *phzH* (blue) form groups with similar gene content. (**b**), (**c**) Box-and-whisker plots showing significantly higher (***p < 0.001) PCN production in strains containing *phzH* and significantly higher (**p < 0.01) 2-HPCA production in strains containing *phzO* in KMB+Fe based on paired t-test. Phenazine production data in (**b**) and (**c**) is taken from the phenotype data shown in Fig. 4.2.

Identifying phenazine biosynthesis hits by mGWAS

We used the hybrid genome assemblies and the phenazine production data to carry out an mGWAS for identification of genetic signatures associated with phenazine production. Phenazine biosynthesis was split into a series of phenotypes pertaining to production of PCA, PCN, and total phenazines in KMB and KMB+Fe for the mGWAS analysis (see Materials and Methods for the complete list of phenotypes). **Table S4.3** shows the number of significant hits (*i.e.*, unique DNA sequences or unitigs) obtained for each phenotype in the mGWAS. All hits and their reverse-complement sequences were aligned to genomes of all strains to find their genomic locations, resulting in a 'preliminary list' of 2493 significant hits across all phenotypes (**Fig. 4.5a**, **Supplementary File 4.1**). Each unitig (and its reverse-complement) in this list may be found in one or more strains and in one or more phenotypes, thus creating redundancies in the list as the unitigs were counted multiple times. These redundancies were eliminated by collapsing the hit list in 3 stages – phenotype-collapsing, strain-collapsing, and reverse-complement collapsing – to result in a 'final list' that only contains a unique entry for each significant hit influencing phenazine biosynthesis. Phenotype collapsing reduced the list to 1568; strain collapsing reduced the list further to 474. Finally, removing entries that were due to reverse complement redundancy resulted in a final list of 330 unique genomic hits for phenazine biosynthesis. The corrected p-value of the hits in the final list is shown in **Fig. 4.5b**. **Fig. S4.3** illustrates the collapsing pipeline, and **Supplementary File 4.2** contains all entries in each of the collapsed lists. A vast majority of the hits (284 out of 330) have a positive effect on phenazine production while the remaining ones had a negative effect (**Fig. S4.4**).

To visualize the genomic location of hits in the final list, we created a circos plot and mapped as many hits as possible to a PCN- or PCA-producing strain (**Fig. 4.5d**). Since each strain contains a different set of accessory genes and many of the hits map to the accessory genome, we were unable to map all hits to a single strain. We selected NCCB 100368[1] as the basis for displaying the PCN-related hits because it contains the majority of the PCN hits (122 out of 127). Similarly, NCCB 88062[1] was selected as the basis to display PCA related hits as it contained 170 out of 203 PCA-related hits, more than any other strain. In total, the two strains combined contain 292 hits from the final list (out of 330). Mapping the hits revealed that there is little association between the PCA hits and PCN-producing strains. In addition, the two hits related to total phenazine production mapped to PCN-producing strains only. In comparison, many of the hits related to PCN biosynthesis were found in both PCA- and PCN-producing strains.

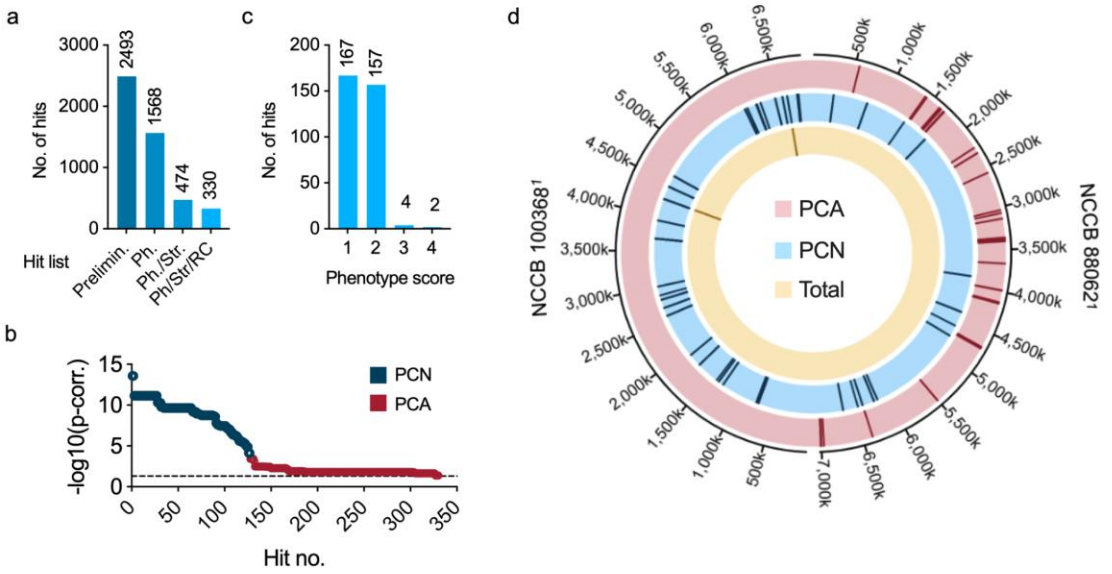


Fig. 4.5. Results of the mGWAS analysis for phenazine production. (**a**) Number of significant mGWAS hits in the preliminary (uncollapsed) list and lists obtained after each collapsing stage - phenotype-collapsed list (Ph), phenotype+strain-collapsed list (Ph./Str.), and phenotype+strain+reverse complement-collapsed list (Ph/Str/RC; also called the 'final list'). Numbers above each bar indicate the exact number of hits in the list corresponding to that bar. (**b**) Corrected p-values of the 330 hits in the final list. Hits were numbered in decreasing order of -log10(p-corrected) value, and were grouped into those influencing PCA production and PCN production. (**c**) Phenotype score distribution of hits in the final list. Numbers above each bar indicate the total number of hits having phenotype score corresponding to that bar. (**d**) Circos plot showing genomic locations of hits in the final list grouped into 3 categories based on the phenotype(s) in which they were significant: PCA production, PCN production, and total phenazine production, with respect to 2 strains - NCCB 100368[1] and NCCB 88062[1].

The mGWAS analysis links unitigs to phenazine production, the next step in our analysis is to identify which genes are associated with each unitig. This mapping is straightforward when the unitig partially or completely overlaps to a CDS, but the connection to a gene is less clear when the unitig is contained within an intergenic region. In these cases, we identified genes upstream and downstream of the intergenic region as potential genes of interest. In total, the 330 unitig hits map to 158 genes in the pangenome (many of the genes were associated with more than one unitig; see **Supplementary File 4.3**). The 158 genes include 80 functionally annotated proteins or homologs of proteins with known function and 78 hypothetical proteins. While the function of the hypothetical proteins are unknown, 33 of them belonged to the core genome while the remainder belonged to the accessory genome. Forty-one out of the 80 proteins of known function belonged to the *P. chlororaphis* core genome.

In the phenotype-collapsing stage we associated each hit to a 'phenotype score', which represents the number of phenotypes in which a hit was found (**Fig. S4.3**); the higher the score, the greater the number of phenotypes influenced by the hit. The phenotype score was used as a metric to identify hits and associated genes most likely to improve phenazine biosynthesis. **Fig. 4.5c** shows the distribution of phenotype scores for the final list. We deemed hits with a phenotype score of 3 or higher as most likely to affect phenazine biosynthesis, and therefore ones that we sought to target for further analysis. The top phenazine-producing strain DSM 21509 primarily produces PCN, and the hits influencing PCN production have lower q-values than those influencing PCA production (Fig. 5b). We therefore narrowed our mGWAS validation and metabolic engineering studies to PCN producing hits only. **Table 4.2** shows the attributes of these hits, two of which have a phenotype score of 4, while the other two have a score of 3. All of these hits were found to be single nucleotide polymorphisms (SNPs) in the coding or intergenic regions. Genes containing or adjacent to SNPs for PCN production include YbhH (a putative isomerase), RhtA (threonine/homoserine exporter), UctC (acetyl-CoA:oxalate CoA-transferase), ProY_1 (proline-specific permease), HutH2 (histidine ammonia-lyase), and two hypothetical proteins (annotated as PS__04251 and PS__04252 in DSM 21509), all of which belong to the *P. chlororaphis* core genome. DSM 21509, the highest PCN-producing strain, contains 3 of the 4 SNPs for PCN production.

Table 4.2. Most influential hits for PCN production, identified in the final list of significant mGWAS hits. Hits are numbered 1 through 4, and attributes such as phenotype score, associated phenotypes, strains, and genes, as well as variant type, p-value, effect on phenotype (positive/negative), & genomic region have been provided for each hit.

No.	Phen. score	Major phen.	Associated gene(s)	Corr. p-value (effect)	Genomic region	Type of variant	Strains
1	4	PCN prodn.	hypothetical protein (PS__04251)	2.6 x 10^{-14} (+)	intergenic	SNP	DSM 21509, DSM 29578^{1}, DSM 29578^{2}, NCCB 100368^{1}
		Total phz. prodn.	hypothetical protein (PS__04252)				
2	4	PCN prodn.	Putative isomerase (YbhH)	2.6 x 10^{-14} (+)	CDS	SNP	DSM 21509, DSM 29578^{1}, DSM 29578^{2}, NCCB 100368^{1}
		Total phz. prodn.					
3	3	PCN prodn.	Threonine/homoserine exporter (RhtA)	6.7 x 10^{-10} (-)	intergenic	SNP	DSM 29578^{1}, DSM 29578^{2}, NCCB 100368^{2}
			Acetyl-CoA:oxalate CoA-transferase (UctC)				
4	3	PCN prodn.	Proline-specific permease (ProY_1)	2.2 x 10^{-9} (+)	intergenic	SNP	DSM 21509, DSM 29578^{1}, DSM 29578^{2}, NCCB 100368^{1}, NCCB 47033, ATCC 17411, ATCC 17809, ATCC 17414, ATCC 9446, ATCC 17814, ATCC 9447, ATCC 17811, ATCC 17810
			Histidine ammonia-lyase (HutH_2)				

Validating mGWAS hits for PCN production

We overexpressed the top gene hits for PCN production in DSM 21509 to verify their phenotypic effects. Three of the genes yielded significant changes in PCN production when overexpressed in DSM 21509. Overexpression of PS__04252 in KMB produced 19.1 ± 3.5 mg/L of PCN compared to 58.7 ± 12.3 mg/L produced by the empty vector control, which is a 67.5% decrease (**Fig. 4.6a**). This target whose overexpression decreases PCN production could provide a target for knockout to increase PCN production. Overexpressing ProY_1 and PS__04251 increased PCN production in KMB+Fe to 420.2 ± 19.7 mg/L and 400.1 ± 21.5 mg/L, respectively, compared to the 343.6 ± 7.3 mg/L PCN produced by the empty vector control (**Fig. 4.6b**). The hits validated here could be combined with other known beneficial genetic manipulations and/or applied to previously engineered and process-optimized strains ([8]: 11.45 g/L PCN) to further improve phenazine production.

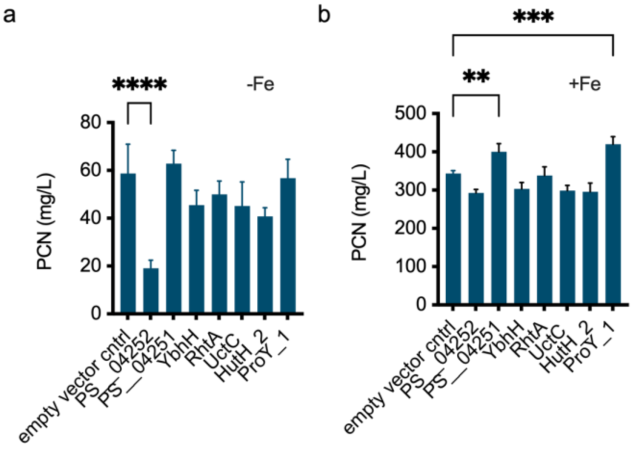


Fig. 4.6. Overexpressing top hits in the top PCN-producer DSM 21509. Genes associated with top phenotype-scoring unitigs for PCN production were overexpressed in the top PCN-producing strain DSM 21509. PCN was quantified after 48 hr of culture in (**a**) King's Media B and (**b**) King's Media B + Fe. Bars represent the average of 3 replicates and error bars represent one standard deviation. Asterisks denote p-values <0.01 (**), <0.001 (***) and <0.0001 (****) when performing an ordinary one-way ANOVA comparison to the empty vector control.

PS__04251 and PS__04252 were both associated with unitig 1. PS__04252 is annotated as a hypothetical protein and has 57% sequence identity and 97% coverage to helix-turn-helix (HTH) MarR-family transcriptional regulator PA1607 from *P. aeruginosa* (NCBI Reference Sequence NP_250298.1) [24]. The MarR, or multiple antibiotic resistance repressor, family often regulates expression of multidrug efflux pumps and some, including PA1607, may derepress in response to oxidative stress [24,25]. Therefore overexpression of the putative regulator PS__04252 could lead to increased repression of its target which could explain the observed decrease, rather than increase, in PCN production. PS__04251 is a hypothetical protein with unknown function which has 86% identity and 100% coverage to putative M14-type zinc cytosolic carboxypeptidase PSF113_3889 (NCBI Reference Sequence WP_041476041.1) from *Pseudomonas ogarae* whose function is unknown [26]. While the function of this and similar proteins are unknown, we found its overexpression to significantly improve phenazine production in *P. chlororaphis*. Because both CDS surrounding unitig 1 were successful, this genomic region could be of further interest to investigate for phenazine production.

The other hit which significantly improved PCN production, ProY_1, was associated with unitig 4 which occurred within the hut operon, which is responsible for histidine catabolism. Due to its position in the highly conserved operon and its sequence similarity, the hit annotated as ProY_1 is likely the histidine permease HutT which imports histidine and is required for its utilization [27]. One study suggests that histidine catabolism in *P. fluorescens* is connected to oxidative stress response, as it could increase intracellular pools of the antioxidant α-ketoglutarate [28]. Possibly, overexpressing HutT could contribute to improved oxidative stress tolerance through a similar mechanism in *P. chlororaphis* by increasing intracellular histidine levels and therefore α-ketoglutarate levels. The other CDS adjacent to unitig 4 is HutH_2 (HutH2), the histidine ammonia lyase which catalyzes the first step of histidine catabolism, the conversion of L-histidine into urocanate [29]. While this hit is associated with the same operon, its overexpression did not affect PCN titers. Because unitig 4 was within the non-coding region between 2 CDS, the mGWAS results in actuality may have been connected to one CDS and not the other. By overexpressing both hits, we were able to identify the one which was relevant for phenazine production.

While overexpressing the genes associated with unitigs 2 and 3 did not produce an observable change in phenotype, they also appear to be related to oxidative stress as well as amino acid export and catabolism. Unitig 2 occurred within the gene annotated as putative isomerase YbhH. This CDS has 34% identity and 92% coverage to *E. coli* YbhH, which does not have a known function but its expression has been upregulated in response to the σ^E stress signaling pathway [30]. Unitig 3 occurred within a non-coding region which was flanked by 16s rRNA and either UctC or RhtA. The hit annotated as UctC has 87% sequence identity and 100% coverage to *P. putida* KT2440 glutarate-CoA transferase GcoT (NCBI Reference Sequence NP_742328.1) which is part of L-lysine catabolism [31]. The other associated hit has 57% identity and 93% coverage to *E. coli* RhtA (NCBI Reference Sequence WP_001295297), which exports threonine and homoserine in addition to other amino acids. While the mGWAS analysis identified these unitigs as significantly related to phenazine production, the overexpression studies showed that they did not alter PCN titers in DSM 21509.

Traditional approaches to metabolic engineering involve manipulation of native or engineered pathways in microbial hosts to direct metabolic flux towards synthesis of a target biochemical [32]. The development of high-throughput technologies, such as next-generation sequencing (NGS), that generate massive amounts of data has made it possible to look for genetic targets spanning the entire genome, thereby facilitating strain engineering for various phenotypes of industrial importance. So far, this population genomics approach for metabolic engineering has been mainly applied to model organisms like *Saccharomyces cerevisiae* for design of strains manifesting higher biochemical production and higher tolerance to growth inhibitors present in the feedstock [33–35]. These studies exploit the natural diversity of strains to identify non-intuitive genetic variants underpinning the trait of interest. In this study, we sought to extend this approach to the production of phenazine compounds in the non-model bacterium *P. chlororaphis*. We used a collection of 34 *Pseudomonas* strains that were isolated from various environmental locations to ensure sufficient genomic diversity. We exploited this diversity within the strain collection to identify 330 variants within the pangenome that influence phenazine production. These pangenome-wide variants were associated to 158 genes, which serve as potential metabolic engineering targets for increasing phenazine biosynthesis in *P. chlororaphis*. To validate our metabolic engineering approach, we selected 7 gene hits and overexpressed them in DSM 21509, the strain with the highest PCN titers. We used the phenotype score to prioritize these hits over others because this method was unbiased in that it is blinded to gene function. Overall this data-driven approach was successful because we identified two candidate genes that improved phenazine production and one that reduced it.

An alternative approach is to prioritize hits based on a rational design strategy, that is, target genes with known functions related to the phenazine biosynthesis or associated pathways. For example, our hit list also includes GacA, a global transcriptional regulator known to impact phenazine production [8,36]. Pursuing this target or others associated with the Gac regulatory cascade could be promising for phenazine production, as the mGWAS and literature results are in agreement. Moreover, since we verified successful hits that may be transcriptional regulators or are involved in oxidative stress response, hits from the list with similar functions could be prioritized for future studies. The other hits on the list which are annotated as HTH-transcriptional regulators (*e.g.*, ArgP, BenM, CynR, MtrA, and RhaS) or the hits known to be connected to oxidative stress response from the literature (*e.g.*, RscC [30], glutathione synthase [37,38] and glucose-6-phosphate 1-dehydrogenase (zwf) [39]) could be pursued as targets in future studies. Importantly, our population genomics approach to metabolic engineering and the hits it generates could be used in tandem with other successful genetic targets from the literature and process optimization strategies to achieve additional improvements in product titers.

Many of the strains in our collection were isolated from diverse environmental locations to ensure natural phenotypic and genomic variation. In our analysis, we observed a broad range of phenotypes and genotypes that were distributed so that both positive and negative groups were well-represented. For example, similar numbers of strains contained *phzH* (14 strains) vs. *phzO* (19 strains), and similar numbers of strains produced no phenazines (10 strains), less than 100 mg/L phenazines (11 strains), and greater than 100 mg/L phenazines (13 strains). This phenotypic diversity, along with the genetic diversity found in the pangenome, allowed us to use a relatively small number of strains to perform the mGWAS analysis and obtain significant hits. The statistical power of hit identification can be potentially improved by using a bigger and more diverse collection of isolates as input to the mGWAS analysis, entailing a larger representation of phenotypic and genomic diversity. Further, GO and pathway enrichment tests of genes associated with 330 hits resulted in no enriched terms (see **Supplementary File 4.4**, **Supplementary File 4.5**). Pursuing hits based off their functions will become more promising as more microbes are annotated for GO-terms and metabolic pathways.

Taken together, this work presents a new approach that enables genome-scale metabolic engineering of pseudomonads. This approach is data-driven, using the power of low cost sequencing and high throughput phenotyping to generate large data sets that correlate desired traits to genomic variants within a microbial population, thereby generating new metabolic engineering targets. While we demonstrate this approach in *Pseudomonas*, it can be extended to other microbial species, especially non-conventional microbes exhibiting industrially relevant phenotypes as new strains are discovered and whole genome sequences become available. Upon gene target elucidation, the microbial strains can be engineered to improve biochemical production or tolerance to various environmental stresses inhibiting cell growth, among other phenotypes. Identification of genetic targets for engineering more complex phenotypes could be accomplished by using a collection of isolates belonging to different but related species, resulting in greater genetic diversity and hence, a more complex pangenome. While rational design strategies for many of these phenotypes may have been previously developed, novel hits identified using a population genomics approach could be used in conjunction with those to further enhance the phenotype of interest and consequently scale-up industrial bioprocesses.

4.4 Conclusions

Advancements in whole genome DNA sequencing and genome-editing techniques, as well as increased availability of bioinformatics tools for analysis of genome-wide data have allowed us to identify metabolic engineering targets spanning the entire pangenome. The accessory genome and core genome are promising sources of metabolic engineering targets for the bacterial production of secondary metabolites such as phenazines. The present study taps into both of these pangenome components to help identify strain engineering targets for biosynthesis of the phenazine PCN in *Pseudomonas chlororaphis*. This pangenome-wide approach, in combination with rational design approaches, could potentially lead to substantial improvement in the phenotype of interest, while also assisting with selection of the appropriate host strain for metabolic engineering.

4.5 Materials and Methods

Strain selection and culturing

All strains designated as *Pseudomonas chlororaphis* that were available as of April and October 2019 were ordered from the American Type Culture Collection (ATCC; Manassas, VA); all strains designated as *Pseudomonas chlororaphis* that were nonredundant and available as of March 2020 were ordered from the German Collection of Microorganisms and Cell Cultures (DSMZ GmbH; Braunschweig, Germany) and the Westerdijk Fungal Biodiversity Institute's Netherlands Culture Collection of Bacteria (NCCB; Utrecht, Netherlands). Strains which appeared to have more than one colony morphology present were separated into distinct isolates (denoted by ^{1} and ^{2}, which were arbitrarily assigned) that were sequenced and cultured separately. All isolates were sequenced with 16s rRNA sequencing (GENEWIZ®; South Plainfield, NJ), and the 33 confirmed *P. chlororaphis* isolates were used in this study (**Table S4.1**). One PCA-producing *P. synxantha* isolate was also used in this study as a phylogenetic outgroup for a total of 34 isolates.

Strains were initially revived according to the guidance of each culture collection then subsequently cultured at 30°C in King's Media B (KMB), the standard media for fluorescent pseudomonads culture, according to the methods of King et al. [40]. To improve phenazine production, KMB+Fe Media was made by supplementing KMB with 100 μM ferric sodium ethylenediaminetetraacetate (FeNaEDTA) based off the findings of van Rij et al. [16]. For phenazine production experiments, both KMB and KMB+Fe contained 1.5 g/L $MgSO_4$. Cultures were supplemented with 50 μg/mL kanamycin sulfate when an antibiotic resistance marker was used. Luria Bertani (LB) broth and TOP10 chemically competent *E. coli* cells were used for cloning.

Liquid culturing was performed using sterile 2 mL 96-deep well plates within an INFORS HT Multitron Pro plate shaker incubator at 1000 rpm and ~88% humidity. Overnight cultures were started by inoculating 500 µL media of interest with the respective colony or glycerol stock. After the overnight culture was incubated with shaking at 30 °C for 22-24 hours, the plate was spun down in Beckman Coulter Allegra 25R centrifuge for 10 minutes at 5,000g. To reduce phenazine transfer and to ensure biofilm-forming strains were well-mixed, old media was removed, and cultures were resuspended in fresh media. To start experimental cultures, 500 µL of desired media were inoculated with 10 µL of resuspended overnight culture. For cultures requiring induction with isopropyl β-D-1-thiogalactopyranoside (IPTG), sterile-filtered IPTG was added to cultures to a final concentration of 1 mM about 4 hours after inoculation.

Phenazine quantification

48 hours after inoculation, phenazine compounds were extracted from each liquid culture using ethyl acetate liquid-liquid extraction. Whole cultures were acidified with 10 µL of 3M HCl, then 1.2 mL ethyl acetate was added to each culture. Each mixture was transferred to a microcentrifuge tube, vortexed at maximum speed for 1 minute, and spun down to separate liquid phases. The ethyl acetate phase was evaporated, resuspended in methanol, and filtered for quantification via HPLC.

All phenazines were quantified with a photodiode-array detector on a Shimadzu Nexera-i LC-2040C 3D liquid chromatograph with an Agilent Poroshell 120 EC-C18 2.7

μm 3.0 x 75 mm column and 3.0 mm x 5.0 mm guard column at 40°C. To resolve the similar phenazine derivatives, the following method with gradients of methanol and ammonium acetate buffer (pH 5.0) was used: 2 μL sample injection, 5 min of 20% methanol, 2 min gradient from 20% to 30% methanol, and 13 min gradient from 30 to 40% methanol with subsequent steps to wash and re-equilibrate the column, with all steps at a 1 mL/min flow rate. PCA and PCN peaks were identified by comparing retention times to those of purchased PCA and PCN (ChemScene; Monmouth Junction, NJ). Because pure 2-HP and 2-HPCA were not commercially available, the identities of these HPLC peaks were confirmed with LC-MS following the same protocol. Phenazines were quantified by converting peak areas at a wavelength of 254 nm and bandwidth of 4 nm to concentrations using extinction constants calculated from the purchased PCA and PCN.

Biofilm formation and growth temperature phenotyping

Biofilm formation phenotyping was characterized using a crystal violet staining assay following the protocol of O'Toole [41]. To adapt the protocol for *Pseudomonas chlororaphis,* a 2% inoculum of overnight culture in the respective media was used to start stationary cultures which were incubated without agitation at 30 °C for 48 hours.

For growth phenotyping, *Pseudomonas* cultures were grown in KMB for 48 hours, diluted to an OD_{600} of 1, then 1 μL was transferred onto respective solid media and incubated at two different temperatures: 30 °C and 37 °C. Four technical replicates were performed for each sample. The plates were imaged every 24 hours for 3 days with an Epson V850 scanner and the images were processed using Iris v0.9.7 (mode: Colony Growth). Any colonies that were missed were reprocessed using Colony Picker in Iris.

Similar to previous studies [42,43], the opacity of colonies was used as an indicator for colony growth. Average opacity for each strain was calculated as the mean of opacity values across all replicates. The value of average opacity on day 2 (*i.e.,* after 48 hours of growth) was used to gauge the ability of strains to grow at the respective temperature levels. Opacity values of all strains in each condition for days 1-3 have been provided in **Supplementary File 4.6**.

Genome assembly

Genomic DNA was isolated using the Quick-DNA™ Fungal/Bacterial Miniprep Kit (Zymo Research; Irvine, CA) and sent to the Microbial Genome Sequencing Center (Pittsburgh, PA) for whole genome sequencing. All isolates were sequenced on the NextSeq 2000 (Illumina; San Diego, CA) with paired-end 150 base pair reads and with Oxford Nanopore technologies. Illumina read quality was assessed using FASTQC v0.11.9 [44] and Nanopore read statistics were assessed with NanoStats v1.28.2 [45] on the Galaxy platform before and after read filtering and trimming (the parameters for all bioinformatics tools are available in **Table S4.2**). Summary statistics (*i.e.,* total assembly length, number of contigs, N_{50}, L_{50}, % GC) were calculated using QUAST v.5.0.2 [46]. Genome completeness was assessed by running BUSCO v.5.2.2 in genome mode using the pseudomonadales_odb10 (prokaryota, 2020-03-06) database [47].

Flye genome assemblies were created with Flye v.2.8.3 and raw Nanopore reads as input [48]. All other genome assemblies used reads which were filtered and trimmed based on read quality. Raw Illumina reads were trimmed to remove adapters and low-quality ends using Trimmomatic v.0.38 [49]. Raw Nanopore reads were adaptor-trimmed using Porechop v.0.2.4 [50] then filtered with filtlong v.0.2.1 [51]. SPAdes genome assemblies were created using SPAdes v.3.12.0 [52] on the Galaxy platform [53]; Unicycler assemblies were created using Unicycler v.0.4.8 using both “Normal” and “Bold” bridging modes and excluding contigs shorter than 1000 bp from the assemblies [54]. The short-reads assemblies were created using only the paired end Trimmomatic output. The long-reads Nanopore assemblies were created using the trimmed and filtered Nanopore reads. The hybrid assemblies were assembled using both sets of aforementioned reads.

Genome annotation and pangenome assembly

Genome assemblies were annotated with Prokka v1.14.6 [55] on the Galaxy platform [53], using a minimum contig size of 1000 and ‘Pseudomonas’ as the genus name. Prokka outputs genome annotations in GFF3 format. These GFF3 files were used along with the draft genome assemblies to generate annotated genome FASTA files by bedtools GetFastaBed v2.30.0 [56]. The pangenome was constructed using PEPPAN v1.0.5 and the .gff files generated by Prokka as input [22]. A rarefaction curve, gene presence absence matrix, and accessory genome tree were created from the PEPPAN output using the included PEPPAN_parser algorithm. Statistics about the core genome were calculated from the gene presence absence matrix using R v4.2.1 (RStudio 2022.07.1). The resulting .nwk tree file was visualized using R v4.2.1 and treeio package v1.20.2 [57]. All remaining figures were created using GraphPad Prism v9.4.1 (GraphPad Software; San Diego, CA).

Genome-wide association study

De novo assembled genomes of the 34 *Pseudomonas* strains were provided as input to DBGWAS v0.5.4 [13] along with the corresponding phenotype values for phenazine production. DBGWAS was implemented for 7 different phenotypes: (i) PCA production in KMB; (ii) PCA production in KMB+Fe; (iii) Effect of Fe on PCA production; (iv) PCN production in KMB; (v) PCN production in KMB+Fe; (vi) Effect of Fe on PCN production; (vii) Total phenazine production in KMB.

For phenotypes (i), (ii), (iv) and (v), concentrations of PCA and PCN (mg/L) obtained from experiments were used directly. Values of phenotypes (iii) and (vi) were estimated by subtracting the concentration of the phenazine compound in KMB from that in KMB+Fe. If the difference was negative, it was replaced by 0. Total phenazine production in KMB was obtained by simply adding the concentrations of all phenazine compounds (*i.e.*, PCA, 2-HPCA, PCN and 2-HP) in KMB. The genome sequences of strains were also provided as BLAST database to DBGWAS for genome mapping of significant unitigs. Significant unitigs were identified based on a corrected p-value cutoff, and a minor allele frequency greater than 1% (default). The number of significant unitigs obtained for each phenotype are listed in **Table S4.3**.

Downstream processing of mGWAS hits

Even though DBGWAS maps significant unitigs to genomes by BLAST, we chose to independently perform unitig alignment to genomes by exact matching to avoid any tolerance to mismatches during alignment. Custom Python3 scripts were used for this purpose with the 34 genome sequences as the mapping database. To ensure that each unitig finds a match, both the significant unitig and its reverse-complement were used. Further, genome annotations were used to determine the genomic regions of the mapped unitigs (*i.e.*, whether the unitig falls within a gene or an intergenic region).

The lists of mGWAS hits from the 7 phenotypes were concatenated into a single list called the 'preliminary list'. In this list, each occurrence of a significant unitig constituted a single entry, creating separate entries for each phenotype, strain, as well as the reverse complement sequence of that unitig. Custom Python3 scripts were used to remove redundancies and collapse the preliminary list so that each significant unitig has a single entry in the final list (**Fig. S4.3**) For each strain, entries for identical unitigs that were significant for multiple phenotypes were collapsed together in the 'phenotype-collapsed' list. Each entry was assigned a phenotype score, which represents the number of phenotypes (out of 7) where each unitig was significant. If a unitig had a phenotype score greater than 1, its corrected p-value was taken to be the minimum of corrected p-values for all phenotypes it is found in. Similarly, the effect of that unitig was taken to be the one with the highest magnitude across all phenotypes. The 'phenotype+strain-collapsed list' combined entries where the same unitig mapped to the same genomic region in multiple strains. Redundancies where the reverse complement of a significant unitig shows up as a separate entry were then collapsed to create the 'final list' of mGWAS hits.

Genes associated to mGWAS hits were determined based on the overlap of unitigs to genes. If the overlap to a gene was partial or complete, that gene was considered to be associated to the unitig. In case of no overlap, i.e., when the unitig appeared completely between two genes, both the neighboring genes were considered to be linked to the unitig.

Experimental validation of top hits

The hits from the final list with a phenotype score of 3 or higher which were significant for PCN production phenotypes were selected as top hits for experimental validation. The CDS immediately upstream and downstream of each significant unitig were chosen as metabolic engineering targets to be overexpressed in the top PCN-producing strain. Any CDS which encoded ribosomal RNA was discarded from the list. If the significant unitig sequence was completely contained within a CDS, only the unitig-containing CDS was studied rather than the 2 adjacent CDS.

Each target was PCR-amplified from the genomic DNA of the strain listed on the top hits file then inserted into the backbone of plasmid pBb(RK2)1k-GFPuv using either restriction digest cloning or NEBuilder® HiFi DNA Assembly (New England Biolabs; Ipswich, MA) using the primers listed in **Table S4.4**. pBb(RK2)1k-GFPuv is a broad-range expression vector with an IPTG-inducible promoter which was gifted by Brian Pfleger at the University of Wisconsin, Madison, and used as the empty vector control [58]. All plasmids used in this study are listed in **Table S4.5**.

Plasmids were transformed into the respective strain via electroporation based on the method of Choi et al. [59]. Electroporations were performed by pulsing either 1.8 or 2.5 kV through a 0.1 or 0.2 cm electroporation cuvette using a MicroPulser™ Electroporation Apparatus (Bio-Rad) then recovering the reaction for 2-3 hours at 30°C.

Culturing and phenazine quantification were performed as described in previous subsections.

Gene Ontology enrichment analysis

To identify enriched GO-terms for significant mGWAS hits, strain DSM 21509 (highest PCN producer) was used as the reference. GO-IDs for this strain were obtained using Blast2GO v6.0.3 [60]. First, Blast2GO was used to map the annotated genome of strain DSM 21509 to proteins in the *P. chlororaphis* protein file (program: blastx; number of blast hits = 5; HSP length cutoff = 50) obtained from NCBI (Taxonomy ID: 587753). Next, the BLAST hits were mapped to GO-identifiers from the database of the Gene Ontology Consortium [61,62]. Lastly, GO mapped hits were annotated (Hit Filter = 2; Filter GO by taxonomy: g-proteobacteria (taxa: 1236,Gammaproteobacteria)) to obtain additional information, such as enzyme codes, enzyme names and InterPro IDs. GO-enrichment test was performed with the obtained GO-IDs using the tool GOEnrichment v2.0.1 on the Galaxy platform [53]. Annotated genes associated to significant unitigs in the final list were provided as the study set. GO annotations from the strain DSM 21509 were provided as the reference set. All enrichment tests were performed using an FDR-corrected p-value cutoff of 0.05 for enrichment.

Pathway enrichment analysis

A list of existing metabolic pathways (and corresponding genes involved) in *P. chlororaphis* strain PA23 was extracted from KEGG PATHWAY database [63] (prefix: pch) and written into a custom pathway database file (.GMT format) using R 4.2.1

(RStudio 2022.07.1). Sequences of all genes associated to significant unitigs in the final list were BLASTed against proteins of the *P. chlororaphis* strain PA23 (obtained from NCBI) using Blast2GO v6.0.3 [60] (program: blastx; number of blast hits = 5; HSP length cutoff = 50) to find homologs. The list of PA23 gene homologs was then used as input along with the custom pathway database file to perform pathway enrichment analysis using the web version of the tool g:Profiler [64].

Data availability

Sequencing reads and assembled genomes for the 34 *Pseudomonas* isolates have been deposited in the NCBI SRA (BioProject ID: PRJNA932460) and NCBI GenBank databases, respectively. NCBI accession numbers for the assembled genomes have been provided in Table 1. Source data for main figures in the study has been provided in **Supplementary File 4.7**. Scripts used for collapsing mGWAS hits have been provided as **Supplementary File 4.8**.

4.6 References

1. Hollas, A. *et al.* A biomimetic high-capacity phenazine-based anolyte for aqueous organic redox flow batteries. *Nature Energy* **3**, 508–514 (2018).

2. Rabaey, K., Boon, N., Höfte, M. & Verstraete, W. Microbial phenazine production enhances electron transfer in biofuel cells. *Environ. Sci. Technol.* **39**, 3401–3408 (2005).

3. Clifford, E. R. *et al.* Phenazines as model low-midpoint potential electron shuttles for photosynthetic bioelectrochemical systems. *Chem. Sci.* **12**, 3328–3338 (2021).

4. Wan, Y., Liu, H., Xian, M. & Huang, W. Biosynthesis and metabolic engineering of 1-hydroxyphenazine in Pseudomonas chlororaphis H18. *Microb. Cell Fact.* **20**, 235 (2021).

5. Guo, S. *et al.* Metabolic Engineering of Pseudomonas chlororaphis for De Novo Production of Iodinin from Glycerol. *ACS Sustainable Chem. Eng.* **10**, 9194–9204 (2022).

6. Yu, J. M., Wang, D., Ries, T. R., Pierson, L. S., 3rd & Pierson, E. A. An upstream sequence modulates phenazine production at the level of transcription and translation in the biological control strain Pseudomonas chlororaphis 30-84. *PLoS One* **13**, e0193063 (2018).

7. Liu, K., Hu, H., Wang, W. & Zhang, X. Genetic engineering of Pseudomonas chlororaphis GP72 for the enhanced production of 2-Hydroxyphenazine. *Microb. Cell Fact.* **15**, 131 (2016).

8. Li, L. *et al.* Metabolic Engineering of Pseudomonas chlororaphis Qlu-1 for the Enhanced Production of Phenazine-1-carboxamide. *J. Agric. Food Chem.* **68**, 14832–14840 (2020).

9. Liu, W.-H. *et al.* Characterization and Engineering of LX24 with High Production of 2-Hydroxyphenazine. *J. Agric. Food Chem.* **69**, 4778–4784 (2021).

10. Lees, J. A. & Bentley, S. D. Bacterial GWAS: not just gilding the lily. *Nature reviews. Microbiology* vol. 14 406 (2016).

11. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, 238 (2016).

12. Lees, J. A. *et al.* Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nat. Commun.* **7**, 12797 (2016).

13. Jaillard, M. *et al.* A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet.* **14**, e1007758 (2018).

14. Mavrodi, D. V., Blankenfeldt, W. & Thomashow, L. S. Phenazine compounds in fluorescent Pseudomonas spp. biosynthesis and regulation. *Annu. Rev. Phytopathol.* **44**, 417–445 (2006).

15. Chin-A-Woeng, T. F. C. *et al.* Biocontrol by Phenazine-1-carboxamide-Producing Pseudomonas chlororaphis PCL1391 of Tomato Root Rot Caused by Fusarium oxysporum f. sp. radicis-lycopersici. *Mol. Plant. Microbe. Interact.* **11**, 1069–1077 (1998).

16. van Rij, E. T., Wesselink, M., Chin-A-Woeng, T. F. C., Bloemberg, G. V. & Lugtenberg, B. J. J. Influence of environmental conditions on the production of phenazine-1-carboxamide by Pseudomonas chlororaphis PCL1391. *Mol. Plant. Microbe. Interact.* **17**, 557–566 (2004).

17. Burr, S. E., Gobeli, S., Kuhnert, P., Goldschmidt-Clermont, E. & Frey, J. Pseudomonas chlororaphis subsp. piscium subsp. nov., isolated from freshwater fish. *Int. J. Syst. Evol. Microbiol.* **60**, 2753–2757 (2010).

18. 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).

19. Conway, H. F. *et al.* Pseudomonas aureofaciens Kluyver and phenazine alpha-carboxylic acid, its characteristic pigment. *J. Bacteriol.* **72**, 412–417 (1956).

20. Haynes, W. C. & Rhodes, L. J. Comparative taxonomy of crystallogenic strains of Pseudomonas aeruginosa and Pseudomon as chlororaphis. *J. Bacteriol.* **84**, 1080–1084 (1962).

21. Benedetti, I., de Lorenzo, V. & Nikel, P. I. Genetic programming of catalytic Pseudomonas putida biofilms for boosting biodegradation of haloalkanes. *Metab. Eng.* **33**, 109–118 (2016).

22. Zhou, Z., Charlesworth, J. & Achtman, M. Accurate reconstruction of bacterial pan- and core genomes with PEPPAN. *Genome Res.* **30**, 1667–1679 (2020).

23. Peix, A. *et al.* Reclassification of Pseudomonas aurantiaca as a synonym of Pseudomonas chlororaphis and proposal of three subspecies, P. chlororaphis subsp. chlororaphis subsp. nov., P. chlororaphis subsp. aureofaciens subsp. nov., comb. nov. and P. chlororaphis subsp. aurantiaca subsp. nov., comb. nov. *Int. J. Syst. Evol. Microbiol.* **57**, 1286–1290 (2007).

24. Kaur, G. & Subramanian, S. The Ku-Mar zinc finger: A segment-swapped zinc ribbon in MarR-like transcription regulators related to the Ku bridge. *J. Struct. Biol.* **191**, 281–289 (2015).

25. Housseini B Issa, K., Phan, G. & Broutin, I. Functional Mechanism of the Efflux Pumps Transcription Regulators From Pseudomonas aeruginosa Based on 3D Structures. *Front Mol Biosci* **5**, 57 (2018).

26. Rimsa, V., Eadsforth, T. C., Joosten, R. P. & Hunter, W. N. High-resolution structure of the M14-type cytosolic carboxypeptidase from Burkholderia cenocepacia refined exploiting PDB_REDO strategies. *Acta Crystallogr. D Biol. Crystallogr.* **70**, 279–289 (2014).

27. Zhang, X.-X. *et al.* Variation in transport explains polymorphism of histidine and urocanate utilization in a natural Pseudomonas population. *Environ. Microbiol.* **14**, 1941–1951 (2012).

28. Lemire, J. *et al.* Histidine is a source of the antioxidant, alpha-ketoglutarate, in Pseudomonas fluorescens challenged by oxidative stress. *FEMS Microbiol. Lett.* **309**, 170–177 (2010).

29. Zhang, X.-X. & Rainey, P. B. Genetic analysis of the histidine utilization (hut) genes in Pseudomonas fluorescens SBW25. *Genetics* **176**, 2165–2176 (2007).

30. Bury-Moné, S. *et al.* Global analysis of extracytoplasmic stress signaling in Escherichia coli. *PLoS Genet.* **5**, e1000651 (2009).

31. Zhang, M. *et al.* Regulation of Glutarate Catabolism by GntR Family Regulator CsiR and LysR Family Regulator GcdR in Pseudomonas putida KT2440. *MBio* **10**, (2019).

32. 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).

33. Hubmann, G. *et al.* Quantitative trait analysis of yeast biodiversity yields novel gene tools for metabolic engineering. *Metab. Eng.* **17**, 68–81 (2013).

34. Maurer, M. J. *et al.* Quantitative Trait Loci (QTL)-Guided Metabolic Engineering of a Complex Trait. *ACS Synth. Biol.* **6**, 566–581 (2017).

35. Meijnen, J.-P. *et al.* Polygenic analysis and targeted improvement of the complex trait of high acetic acid tolerance in the yeast Saccharomyces cerevisiae. *Biotechnol. Biofuels* **9**, 5 (2016).

36. Wang, D. *et al.* Roles of the Gac-Rsm pathway in the regulation of phenazine biosynthesis in Pseudomonas chlororaphis 30-84. *Microbiologyopen* **2**, 505–524 (2013).

37. Nikel, P. I. *et al.* Reconfiguration of metabolic fluxes in Pseudomonas putida as a response to sub-lethal oxidative stress. *ISME J.* **15**, 1751–1766 (2021).

38. Wongsaroj, L. *et al.* Pseudomonas aeruginosa glutathione biosynthesis genes play multiple roles in stress protection, bacterial virulence and biofilm formation. *PLoS One* **13**, e0205815 (2018).

39. Kim, J., Jeon, C. O. & Park, W. Dual regulation of zwf-1 by both 2-keto-3-deoxy-6-phosphogluconate and oxidative stress in Pseudomonas putida. *Microbiology* **154**, 3905–3916 (2008).

40. King, E. O., Ward, M. K. & Raney, D. E. Two simple media for the demonstration of pyocyanin and fluorescin. *J. Lab. Clin. Med.* **44**, 301–307 (1954).

41. O'Toole, G. A. Microtiter dish biofilm formation assay. *J. Vis. Exp.* (2011) doi:10.3791/2437.

42. Banzhaf, M. *et al.* Outer membrane lipoprotein NlpI scaffolds peptidoglycan hydrolases within multi-enzyme complexes in Escherichia coli. *EMBO J.* **39**, e102246 (2020).

43. Ropars, J. *et al.* Domestication of the Emblematic White Cheese-Making Fungus Penicillium camemberti and Its Diversification into Two Varieties. *Curr. Biol.* **30**, 4441–4453.e4 (2020).

44. Andrews, S. & Others. FastQC: a quality control tool for high throughput sequence data. Preprint at (2010).

45. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

46. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. & Gurevich, A. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**, i142–i150 (2018).

47. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).

48. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

49. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

50. Wick, R. R. Porechop: adapter trimmer for Oxford Nanopore reads. Preprint at (2018).

51. Wick, R. & Menzel, P. Filtlong: quality filtering tool for long reads. Preprint at (2019).

52. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinformatics* **70**, e102 (2020).

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

54. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).

55. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

57. Wang, L.-G. *et al.* Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol. Biol. Evol.* **37**, 599–603 (2020).

58. Cook, T. B. *et al.* Genetic tools for reliable gene expression and recombineering in Pseudomonas putida. *J. Ind. Microbiol. Biotechnol.* **45**, 517–527 (2018).

59. Choi, K.-H., Kumar, A. & Schweizer, H. P. A 10-min method for preparation of highly electrocompetent Pseudomonas aeruginosa cells: application for DNA fragment transfer between chromosomes and plasmid transformation. *J. Microbiol. Methods* **64**, 391–397 (2006).

60. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).

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

62. Harris, M. A. *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**, D258–61 (2004).

63. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

64. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).

4.7 Supplementary Information

Supplementary Figures

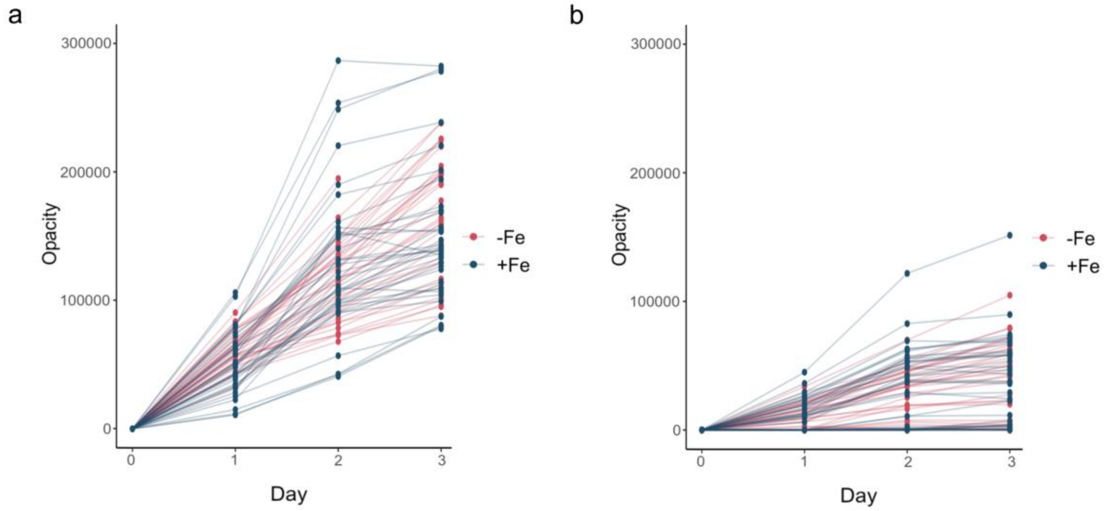


Fig. S4.1. Growth curves of Pseudomonas strains at two different temperatures. Growth curves of the 34 Pseudomonas strains based on opacity values for the first 3 days of growth at (**a**) 30 °C, and (**b**) 37 °C in two media conditions – KMB (pink) and KMB+Fe (blue). Data points represent average opacity for a particular strain on a given day and condition.

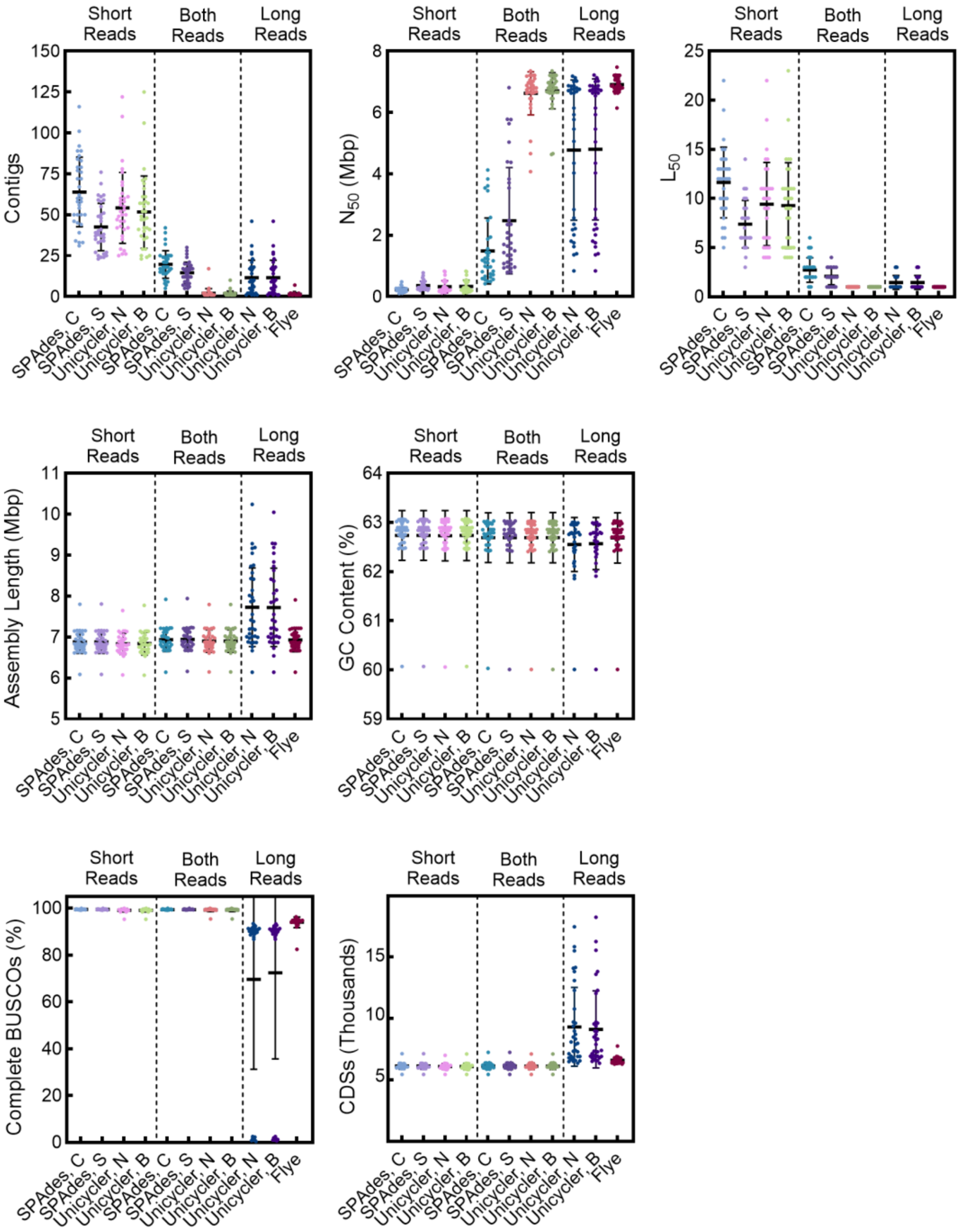


Fig. S4.2. Graphical comparisons of the summary statistics for all genome assemblies. Genome assemblies were created with only short Illumina sequencing reads (left graph subsections), only long Oxford Nanopore sequencing reads (right graph subsection) or both reads sets in a hybrid approach (middle graph subsection), with contigs files (C) and scaffolds (S) generated by SPAdes, and genome assemblies generated by using either normal (N) or bold (B) bridging mode in Unicycler. Number of contigs, N_{50}, L_{50}, total sequence length, and GC content were generated using QUAST. The % complete BUSCOS was calculated using BUSCO and number of CDS from assembly annotations generated by Prokka.

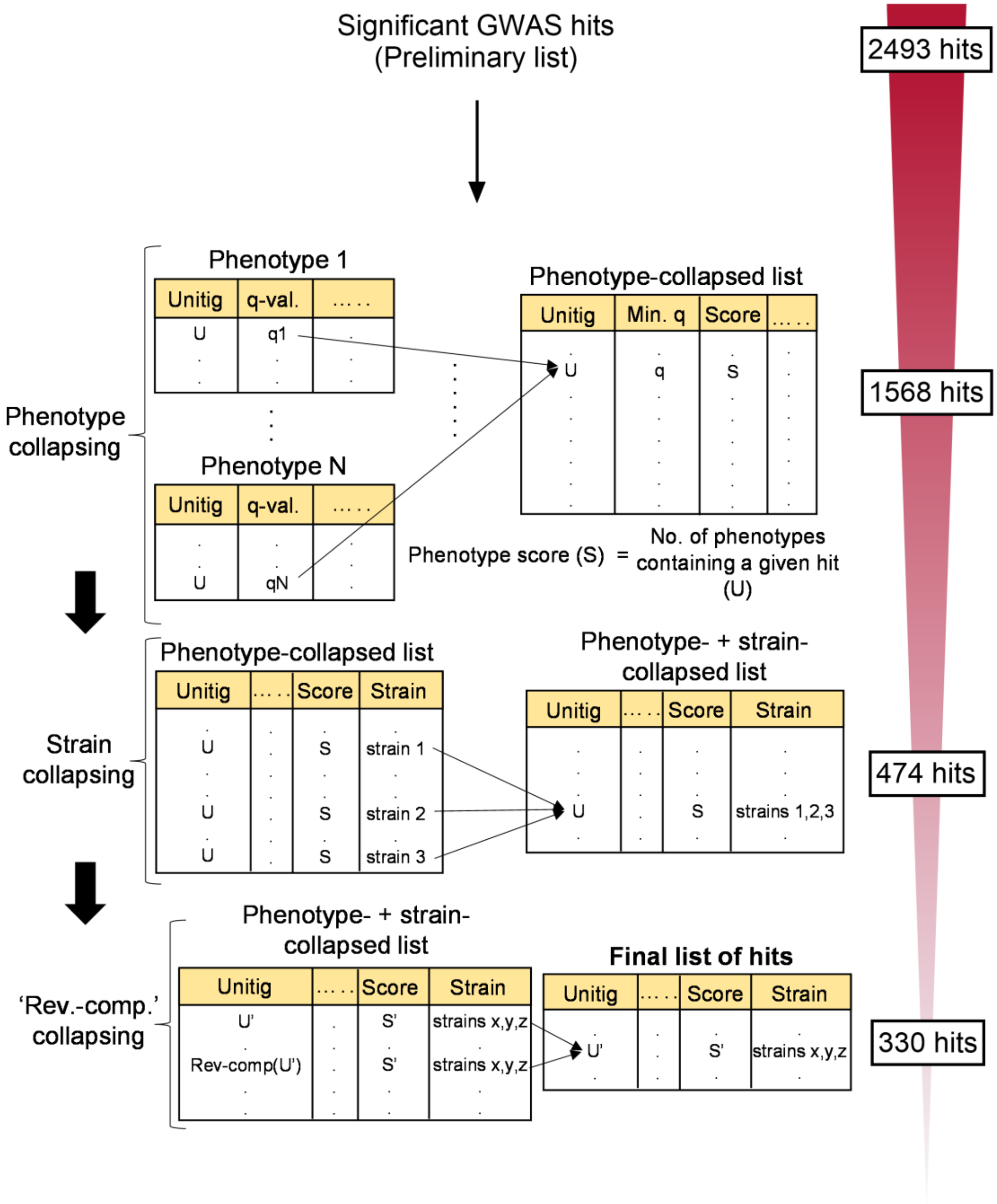


Fig. S4.3. Collapsing pipeline for GWAS hits. A schematic showing each collapsing stage to remove redundancies in the preliminary list of significant GWAS hits. A phenotype score is assigned to each hit in the phenotype-collapsing stage. Numbers on the right indicate the total number of hits in the preliminary list as well as resulting lists from each corresponding collapsing stage for the phenazine production GWAS results.

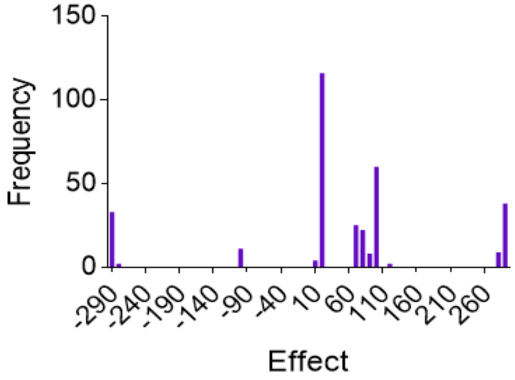


Fig. S4.4. Distribution of significant GWAS hit effects in the 'final list'. Vertical axis indicates the number of significant hits having effect values in a certain range as specified by each bin.

Supplementary Tables

Table S4.1. List of strains used in this study. All of the listed isolates were 16s rRNA sequenced and confirmed to be *Pseudomonas chlororaphis* except for ATCC 17413 which was determined to be *Pseudomonas synxantha*.

Full strain name	Strain isolate used in this study	Source
Pseudomonas chlororaphis subsp. chlororaphis (ATCC® 9446™)	ATCC 9446	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 9447™)	ATCC 9447	American Type Culture Collection
Pseudomonas chlororaphis subsp. aureofaciens (ATCC® 13985™)	ATCC 13985	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 13986™)	ATCC 13986[1] ATCC 13986[2]	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 15926™)	ATCC 15926	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17411™)	ATCC 17411	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17413™)	ATCC 17413	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17414™)	ATCC 17414	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17415™)	ATCC 17415[1] ATCC 17415[2]	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17417™)	ATCC 17417	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17418™)	ATCC 17418[1] ATCC 17418[2]	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17419™)	ATCC 17419	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17809™)	ATCC 17809	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17810™)	ATCC 17810	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17811™)	ATCC 17811	American Type Culture Collection
Pseudomonas chlororaphis (ATCC® 17814™)	ATCC 17814	American Type Culture Collection
Pseudomonas chlororaphis subsp. aurantiaca (ATCC® 33663™)	ATCC 33663[1] ATCC 33663[2]	American Type Culture Collection
Pseudomonas chlororaphis subsp. aureofaciens DSM-6508	DSM 6508	German Collection of Microorganisms and Cell Cultures
Pseudomonas chlororaphis subsp. piscium DSM-21509	DSM 21509	German Collection of Microorganisms and Cell Cultures
Pseudomonas chlororaphis subsp. aureofaciens DSM-29578	DSM 29578[1] DSM 29578[2]	German Collection of Microorganisms and Cell Cultures
Pseudomonas chlororaphis subsp.	NCCB 47033	Netherlands Culture Collection of

chlororaphis NCCB 47033		Bacteria
Pseudomonas chlororaphis subsp. *aureofaciens* NCCB 60037	NCCB 60037	Netherlands Culture Collection of Bacteria
Pseudomonas chlororaphis subsp. *aureofaciens* NCCB 60038	NCCB 60038	Netherlands Culture Collection of Bacteria
Pseudomonas chlororaphis subsp. *aureofaciens* NCCB 82053	NCCB 82053[1] NCCB 82053[2]	Netherlands Culture Collection of Bacteria
Pseudomonas chlororaphis subsp. *aureofaciens* NCCB 88062	NCCB 88062[1] NCCB 88062[2]	Netherlands Culture Collection of Bacteria
Pseudomonas chlororaphis subsp. *chlororaphis* NCCB 100368	NCCB 100368[1] NCCB 100368[2]	Netherlands Culture of Collection Bacteria

Table S4.2. List of bioinformatics tools and parameters used for genome and pangenome assembly and annotation in this study. Other parameters which are not listed were kept at default values.

Tool Name	Version	Parameters
FASTQC	v0.11.9	Input: Illumina reads, raw and output from Trimmomatic (forward and reverse fastq.gz files) ● Default settings
NanoStats	v1.28.2	Input: Oxford Nanopore reads, raw and output from Porechop and filtlong (fastq.gz files) ● Default settings
QUAST	v5.0.2	Input: All genome assemblies (fasta files) ● Default settings
BUSCO	v5.2.2	Input: All genome assemblies (fasta files) ● Lineage dataset (--lineage-dataset): pseudomonadales_odb10 (prokaryota, 2020-03-06) ● Running mode (--mode): genome
Flye	v2.8.3	Input: raw Oxford Nanopore reads (fastq.gz files) ● Specify raw Oxford Nanopore reads as input (--nano-raw)
Trimmomatic	v0.38	Input: raw Illumina reads (forward and reverse fastq.gz files) ● ILLUMINACLIP: NexteraPE ● LEADING:3 ● TRAILING:3 ● SLIDINGWINDOW: 4:15 ● MINLEN:36
Porechop	v0.2.4	Input: raw Oxford Nanopore reads (fastq.gz files) ● Default settings
filtlong	v0.2.1	Input: adapter-trimmed Oxford Nanopore reads output from Porechop (fastq.gz files) ● Minimum length (--min_length): 1000 ● Minimum mean quality (--min_mean_q): 10
SPAdes	v3.12.0	Input: paired-end trimmed Illumina reads output from Trimmomatic (forward and reverse fastq.gz files). Hybrid assemblies also used filtered long reads output from filtlong (fastq.gz files) ● Default settings
Unicycler	v0.4.8	Input: paired-end trimmed reads output from Trimmomatic. Hybrid assemblies also used filtered long reads output from filtlong. ● Bridging mode (--mode): normal OR bold ● Exclude contigs shorter than this length from FASTA (bp) (--min_fasta_length: 1000
Prokka	v1.14.6	Input: All genome assemblies (fasta files) ● Default settings
PEPPAN	v1.0.5	Genome annotations output from Prokka (.gff files for either all final genome assemblies or only the *P. chlororaphis* assemblies)
PEPPAN_parser		Output by PEPPAN output (PEPPAN.gff file) ● Generate gene presence/absence tree (--tree) ● Generate rarefaction curve (--curve) ● Ignore pseudogenes in analysis (--pseudogene). This flag was used in the pangenome analysis, but left of when determining which hits belonged to the core and

		accessory genome
treeio	v1.20.2	Input: gene presence/absence tree output by PEPPAN_parser (All 34 strains, excluding pseudogenes; CDS_content.nwk file)

Table S4.3. Number of significant unitigs obtained for different phenotypes in the GWAS analysis.

Phenotype	No. of significant unitigs
PCA production in KMB	121
PCA production in KMB + Fe	77
Effect of Fe on PCA production	79
PCN production in KMB	122
PCN production in KMB + Fe	81
Effect of Fe on PCN production	4
Total phenazine production in KMB	2

Table S4.4. List of primers used in this study.

Name	Sequence (5'→3')	Description
PS__04252_F	AATTTCAGAATTCAAAAGATCTTTTAAGAAGGAGATATACATATGGTCAAACGCACAAGC	Amplify PS__04252 from DSM 21509 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, forward primer
PS__04252_R	CCTTACTCGAGTTTGGATCCCTATTGCACCGGCACCC	Amplify PS__04252 from DSM 21509 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, reverse primer
PS__04251_F	AATTTCAGAATTCAAAAGATCTTTTAAGAAGGAGATATACATATGACCGTGGCTCAAAGC	Amplify PS__04251 from DSM 21509 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, forward primer
PS__04251_R	ATCCTTACTCGAGTTTGGATCCTCAGCGCAGGATGCCGA	Amplify PS__04251 from DSM 21509 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, reverse primer
RhtA_F	AATTTCAGAATTCAAAAGATCTTTTAAGAAGGAGATATACATATGAATGACCAGCCCCG	Amplify *rhtA* from DSM 29578[2] gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, forward primer
RhtA_R	ATCCTTACTCGAGTTTGGATCCTCAATCAGCTGCAACCAAAG	Amplify *rhtA* from DSM 29578[2] gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, reverse primer
ProY1_F	AATTTCAGAATTCAAAAGATCTTTTAAGAAGGAGATATACATATGCAACAGCAAGCTCAA	Amplify *proY_1* from ATCC 9447 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, forward primer
ProY1_R	ATCCTTACTCGAGTTTGGATCCTTATCGATGGGACAAAGAAGG	Amplify *proY_1* from ATCC 9447 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, reverse primer
UctC_F	AATTTCAGAATTCAAAAGATCTTTTAAGAAGGAGATATACATATGGGCGCGTTATCTCAT	Amplify *uctC* from DSM 29578[2] gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, forward primer
UctC_R	ATCCTTACTCGAGTTTGGATCCTCACAGCACGCCCGAG	Amplify *uctC* from DSM 29578[2] gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, reverse primer
HutH2_F	AATTTCAGAATTCAAAAGATCTTTTAAGAAGGAGATATACATGTGACTGCGCTAAATCTG	Amplify *hutH2* from ATCC 9447 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, forward primer
HutH2_R	ATCCTTACTCGAGTTTGGATCCTTACAGGCTCGGCAGC	Amplify *hutH2* from ATCC 9447 gDNA for restriction digest cloning into pBb(RK2)1k-GFPuv backbone, reverse primer
pBb(RK2)1k_F	GGATCCAAACTCGAGTAAG	Amplify pBb(RK2)1k-GFPuv backbone

		for HiFi assembly, forward primer
pBb(RK2)1k_R	ATGTATATCTCCTTCTTAAAAGATCT	Amplify pBb(RK2)1k-GFPuv backbone for HiFi assembly, reverse primer
YbhH-HiFi_F	TTTAAGAAGGAGATATACATATGTCTTTTGAACTGGACCTTCCC	Amplify *ybhH* from DSM 21509 gDNA for HiFi assembly, forward primer
YbhH-HiFi_R	CCTTACTCGAGTTTGGATCCTTAGCCCCGCCCTTTCAAC	Amplify *ybhH* from DSM 21509 gDNA for HiFi assembly, reverse primer
Seq-pBb(Rk2)1k_F	CAATTAATCATCCGGCTCG	Forward sequencing primer for overexpression plasmids
Seq-pBb(Rk2)1k_R	GACTCTAGTAGAGAGCGTTC	Reverse sequencing primer for overexpression plasmids

Table S4.5. List of plasmids used in this study.

Plasmid name	Description	Source
pBb(RK2)1k-GFPuv	IPTG-inducible trc promoter expressing *gfpuv*, kanamycin resistance, RK2 origin of replication	Cook et al. 2018
pBb(RK2)1k-PS__04252	IPTG-inducible trc promoter expressing PS__04252 amplified from DSM 21509, kanamycin resistance, RK2 origin of replication	This study
pBb(RK2)1k-PS__04251	IPTG-inducible trc promoter expressing PS__04251 amplified from DSM 21509, kanamycin resistance, RK2 origin of replication	This study
pBb(RK2)1k-YbhH	IPTG-inducible trc promoter expressing *ybhH* amplified from DSM 21509, kanamycin resistance, RK2 origin of replication	This study
pBb(RK2)1k-RhtA	IPTG-inducible trc promoter expressing *rhtA* amplified from DSM 29578[2], kanamycin resistance, RK2 origin of replication	This study
pBb(RK2)1k-UctC	IPTG-inducible trc promoter expressing *uctC* from amplified from DSM 29578[2], kanamycin resistance, RK2 origin of replication	This study
pBb(RK2)1k-HutH2	IPTG-inducible trc promoter expressing *hutH2* amplified from ATCC 9447, kanamycin resistance, RK2 origin of replication	This study
pBb(RK2)1k-ProY1	IPTG-inducible trc promoter expressing *proY_1* amplified from ATCC 9447, kanamycin resistance, RK2 origin of replication	This study

Chapter 5: Conclusion and future directions

Microorganisms natively exhibiting industrially relevant phenotypes present attractive hosts for biochemical production. Advancements in synthetic biology and sequencing techniques have facilitated genome-wide manipulation of these microbes in high-throughput screens, allowing us to identify strain engineering targets at a pangenome-scale via bioinformatic analysis of the screening data. This dissertation presents different genome-scale approaches and the corresponding bioinformatic analyses that could be implemented to elucidate previously unknown gene targets for metabolic engineering of microbes in an effort to improve the scalability and economic feasibility of industrial bioprocesses.

Chapter 2 presents an end-to-end pipeline, acCRISPR, that identifies statistically significant genes for a phenotype using data from pooled CRISPR knockout screens. While existing methods for CRISPR screen analysis have been successful in identifying essential genes for growth of mammalian cancer cell lines, they were found to perform poorly on microbial datasets. These methods make use of data from multiple screens and apply Bayesian approaches to predict sgRNA activity. acCRISPR, on the other hand, utilizes experimental activity profiles from the host organism to remove low-activity sgRNA and accurately call essential genes from a single screening dataset, thereby reducing the scale of the screening experiment and presenting a significant step forward in CRISPR screen analysis. The essential gene hits and salt tolerance hits identified by acCRISPR in the commercially important yeast *Y. lipolytica* would be useful to engineer Yarrowia for cultivation in high salt conditions in industrial bioreactors.

Functional genomic screens conducted using CRISPRa and CRISPRi systems have shown success in unraveling the biological mechanisms underlying various phenotypes by modulating expression of genes rather than knocking them out. Having demonstrated the ability to reliably call gene hits from CRISPR knockout screens, acCRISPR could be implemented on CRISPRi and CRISPRa screening datasets, which would help improve our understanding of gene function and relevant biological pathways in non-conventional microbes, such as those involved in cellular stress response, potentially complementing findings from knockout studies. A critical challenge, however, in analyzing CRISPRa/i datasets would be the lack of availability of sgRNA activity profiles for gene perturbation experiments. A potential workaround for this issue could be to use knockout activity profiles for the same guides as a substitute when implementing acCRISPR, given that both gene modulation and gene knockout entail binding of the sgRNA to the target genomic locus, independent of the nature of Cas protein activity (gene disruption/modulation).

Another compelling direction in this area could be to explore the cross-species applicability of acCRISPR. Using existing data for CRISPR screens performed in mammalian cell lines with libraries such as GeCKO, Avana and Yusa, coupled with *in silico* activity scores for library guides, the performance of acCRISPR in accurately predicting the gold standard set of essential genes for tumor cell survival could be evaluated. Furthermore, the accuracy and cross-species applicability of the method could be potentially improved by making the nature of the null distribution more adaptable to the dataset being analyzed. While the assumption of normality for the population of non-essential genes seems to be a reasonable one, it may not always lead to optimum performance for every dataset, given the biological differences among various cell types. The method could instead utilize a more generic distribution function to describe the distribution curve that could resemble for instance, a normal distribution, a t-distribution (typically narrower than normal distribution, with heavier tails), or a skewed t-distribution (asymmetric distribution), depending on the gene fitness profiles. This could be achieved by incorporating hyperparameters into the distribution function, whose values depend on the screening dataset, and ultimately govern the nature of the null distribution.

The *Y. lipolytica* sgRNA activity profiles used for acCRISPR analysis were also employed to examine the sgRNA activity prediction accuracy of deep learning models in Chapter 3. By training a convolutional neural network (CNN) model and a large language model (LLM) on balanced, imbalanced and re-balanced datasets, we found the models to have maximum accuracy in predicting high- and low-activity sgRNA when trained on inherently balanced datasets. Similar performances of the CNN and LLM architectures on the same (CRISPR-Cas12a) training set, and variable performance of the same (CNN) architecture on training sets with different compositions signify the importance of training set characteristics in influencing prediction power. The *Y. lipolytica* CRISPR-Cas12a library that resulted in the best prediction performance consists of substantial fractions of high- and low-activity sgRNA, and was designed using minimal design criteria. This approach of using minimal design criteria could be extended to other species to generate innately balanced training datasets and consequently improve sgRNA design for those species.

While deep learning frameworks such as CNN architectures have shown to be effective in predicting sgRNA activity, they need to be trained from the ground up for every individual species, requiring greater use of computational resources. LLMs, on the other hand, allow the creation of a generalizable pre-trained model that can then be fine-tuned on a wide range of tasks, including sgRNA activity prediction, across multiple species. The ease of usage of AI models to generate species-specific sgRNA activity predictions could thus be improved by pre-training an LLM architecture on a large collection of genomes from species belonging to one or more phylogenetic groups (such as prokaryotes, fungi, plants and/or mammals), and fine-tuning this pre-trained model on CRISPR sgRNA datasets for a particular species of interest. Using the LLM approach would greatly improve the algorithmic efficiency since the computationally intensive pre-training step is a one-time process, while fine-tuning, required to be performed separately for each species, is relatively less burdensome, allowing researchers with limited computational resources to conveniently use the pre-trained model for their specific needs.

Chapter 4 showcases the use of a population genomics approach in enabling identification of 330 pangenome-wide biomarkers influencing phenazine biosynthesis in the bacterium *Pseudomonas chlororaphis*. This data-driven approach resulted in unbiased identification of gene candidates belonging to core and accessory genomes compared to strategies that mainly focus on genes having known relevant functions or genes pertaining to relevant native metabolic pathways. Given the industrial importance of phenazine compounds as bio-fungicides for crop protection, the hits elucidated in this study could be potentially used to produce higher titers of phenazines with *P.chlororaphis* in industrial bioreactors.

While the hits identified in this work add to the existing knowledge of strain engineering targets for improving phenazine production, additional studies could be conducted to further investigate the biological significance of these hits. For instance, the hypothetical protein PS_04252, reducing cellular PCN production upon overexpression in hit validation experiments, was putatively linked to a family of transcriptional regulators. To unravel the function of this gene, transcriptomic studies such as an RNA-seq analysis could be performed to identify genes that are differentially expressed between the wildtype strain and the strain containing overexpression of the hypothetical protein. Similarly, another validated gene hit, ProY_1, was linked to histidine catabolism and increasing intracellular histidine levels. The putative function of this gene could be corroborated by conducting metabolomic studies and measuring titers of histidine and other precursors in the histidine catabolic pathway in a ProY_1 overexpressing strain relative to the wildtype strain background. Follow-up studies like these would not only allow us to establish the role of gene candidates in influencing phenazine synthesis, but also expand our current understanding of *P. chlororaphis* as a biotechnology host.

In addition to a deeper investigation of the identified gene hits, refining the GWAS analysis promises to be another future direction. The GWAS performed in this study used a relatively small collection of 34 isolates to call statistically significant variants for phenazine biosynthesis. This small sample size of strains was due to the limited availability of *P. chlororaphis* isolates at the time this study was conducted. However, as new strains get discovered and their genome sequences become more readily available, the study could be expanded to a larger collection of *P. chlororaphis* isolates, which in turn would improve the statistical power of the GWAS analysis. Moreover, the GWAS method used in this study, DBGWAS, correlates the presence/absence of unitigs (high-confidence contigs obtained my collapsing successive genomic k-mers in a de Bruijn graph) to phenotypic levels to enable hit identification. A major drawback of using the presence/absence information of unitigs in the GWAS is that it fails to detect phenotypic associations for copy number variants (resulting from repeat regions in the genome). Correlating unitig counts to phenotype instead would help overcome this limitation and improve variant detection power of the GWAS analysis.

Taken together, the forward genetic frameworks described in this dissertation present meaningful advancements in enabling identification of genome-wide strain engineering targets and evolving microbes as chemical factories for industrial bioproduction.