Title

Functional Genomic Screening in the Methylotrophic Yeast Komagataella Phaffii

Author

Tafrishi, Aida

Publication Date

2024

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Functional Genomic Screening in the Methylotrophic Yeast *Komagataella phaffii*

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

Doctor of Philosophy

in

Chemical and Environmental Engineering
by

Aida Tafrishi

September 2024

Dissertation Committee:
Dr. Ian Wheeldon, Chairperson
Dr. Robert Jinkerson
Dr. Jason Stajich

The Dissertation of Aida Tafrishi is approved:

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

First and foremost, I am deeply indebted to my advisor, Professor Ian Wheeldon, whose guidance, patience, and generosity have been instrumental in the completion of this dissertation. Your insights and expertise have shaped my work in ways I could never have imagined, and your support has been unwavering. Thank you for your mentorship, for challenging me to think critically, and for always encouraging me to strive for excellence. Your belief in my abilities has been a great source of motivation, and I am incredibly fortunate to have had the opportunity to learn from you.

A special thank you goes to my husband, the love of my life, and my best friend, Farzin, whose patience, love, and understanding have been a true pillar of strength. You have been my biggest supporter, always encouraging me to push through the tough times and celebrating every milestone with me. Your unwavering belief in me, even when I doubted myself, has been the fuel that kept me going. I am so grateful to have you by my side, now and always.

I would like to also express my deepest gratitude to my parents, who have been a constant source of love, support, and encouragement throughout this journey. To my parents, your unwavering belief in me has been my anchor during the most challenging times. Your sacrifices and guidance have shaped who I am today, and for that, I am eternally grateful.

To my brother and sister, Parsa and Bita, thank you for always being there with words of wisdom and for understanding the demands of this journey. Your support, humor, and

belief in my abilities have been invaluable. I am so lucky to have you both by my side, cheering me on every step of the way.

To my friends, you have been my sounding boards, my confidants, and my escape when I needed it most. Thank you for the countless laughs, the late-night conversations, and for being my chosen family. Your friendship has been a vital part of my life, and I couldn’t have done this without you.

Finally, to everyone who has contributed to this journey, whether through advice, support, or simply by being there, thank you. This dissertation is not just a reflection of my work, but a testament to the love, support, and guidance of all the wonderful people in my life.

ABSTRACT OF THE DISSERTATION

Functional Genomic Screening in the Methylotrophic Yeast *Komagataella phaffii*

by

Aida Tafrishi

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

The production of recombinant products, including enzymes, biomaterials, and therapeutics, is a driving force in biotechnology addressing various global challenges. Non-conventional microbes are particularly appealing for metabolic engineering due to their uniquely evolved characteristics that can simplify the engineering process compared to more traditional model organisms. The methylotrophic yeast *Komagataella phaffii* stands out for its ability to grow to high cell densities, perform post-translational modifications, and secrete high titers of recombinant proteins with minimal endogenous host protein secretion. Although significant progress has been made in developing *K. phaffii* to produce biopharmaceuticals and other value-added products, there remains a need for advanced synthetic biology tools facilitating genome engineering, functional genomic screening, and rapid strain optimization to fully harness its potential.

We have sought to overcome these limitations by providing a detailed protocol for designing a highly active genome-wide sgRNA knockout library. We measured the cutting efficiency of the library using an experimental workflow involved with transforming the library into cells with a deficient dominant DNA repair pathway and performing growth screens. Our results demonstrated that over 98% of the sgRNAs in the library were active. The activity validation ensures accurate and precise screening outcomes. We then performed growth screens using glucose as the sole carbon source and defined a set of consensus essential genes for *K. phaffii*. Comparative analysis of these genes with essential genes from other known yeast species revealed a core set of essential genes in *K. phaffii*, many of which are linked to vital cell processes. This analysis also revealed essential genes exclusive to *K. phaffii* related to key metabolic engineering targets, such as protein production, secretion, and glycosylation.

Additionally, we employed ribosome profiling and next-generation sequencing to examine the global and early secretory demands of *K. phaffii*, focusing on host protein synthesis and endoplasmic reticulum trafficking before and after methanol induction. This analysis was conducted using an industrial strain of *K. phaffii* engineered to produce human serum albumin (HSA) under methanol conditions. By identifying key host proteins that impose the greatest constraints on the biogenetic machinery and subsequently targeting these genes using the CRISPR-Cas9 system, we achieved a 35% increase in HSA secretion. The highly active, genome-wide CRISPR library as well as the generated Ribo-req protocol and data developed in this study facilitates functional genomic screening in *K. phaffii*, provides insights into this cell’s biology, and holds potential for enabling a wide range of engineering into this host cell.

Table of Contents

List of figures

List of Tables

Chapter 1: Introduction

1-1 Background

Biotechnology plays a crucial role in addressing some of the most critical challenges of our time, including health, climate change, energy, and food security. The market for biologics is projected to reach to 400 Billion USD by 2025[1]. The COVID-19 pandemic highlighted the vital importance of biotechnology in developing life-saving diagnostics, therapeutics, and vaccines. The growing demand for sustainable and rapid production of biopharmaceuticals has driven significant advancements in biotechnology, particularly in the engineering and production of recombinant proteins. These engineered proteins, ranging from enzymes and structural proteins to therapeutic proteins, have become indispensable in various industries. As of 2015, there were more than 400 marketed recombinant peptides and proteins as well as an additional 1300 undergoing clinical trials[2–4].

The production of recombinant proteins relies on cellular factories like bacteria (E.g., *Escherichia coli*), yeast (E.g., *Saccharomyces cerevisiae*), or mammalian cells (E.g., Chinese Hamster Ovary). The selection of the host relies on various factors including recombinant protein structural complexity, codon preferences, glycosylation requirements, and required minimum titers. Advances in synthetic biology and metabolic engineering have significantly improved our ability to efficiently engineer these cells, enabling the development of customized cellular systems tailored to produce high-quality proteins that meet specific production demands.

Recently, there has been increasing interest in the biotechnology industry in utilizing alternative microbes as microbial factories. While mammalian cells are still widely used in the bio-industry (for example to produce antibodies), alternative microbial hosts offer various distinct advantages. These microbes, particularly eukaryotic organisms such as yeast, are characterized by smaller and more stable genomes, a low risk of contaminating the final product with host cell proteins due to their simpler secretomes, the ability to perform post-translational modifications similar to those of higher eukaryotes, high secretion capacity, and a lack of susceptibility to infectious agents like viruses. Yeasts exhibit faster growth rates compared to mammalian cells, which can lead to shorter production cycles, increased efficiency, and lower product costs. These host cells can tolerate extreme environmental conditions, such as variations in temperature, pH, and osmolarity, making them robust production systems in diverse industrial settings. Additionally, their ability to utilize a wide range of carbon sources further enhances their versatility, facilitating the conversion of inexpensive or alternative feedstocks into added value bioproducts. These attributes make yeasts not only cost-effective but also highly efficient platforms for the large-scale production of bioproducts, positioning them as a promising alternative to traditional mammalian cell systems in the rapidly evolving biotechnology industry[5–7].

Komagataella phaffii, previously known as *Pichia pastoris*, is widely used as a heterologous protein production host[8–13]. This microorganism can grow to high cell densities, is amenable to fast and relatively straightforward genetic engineering, performs post-translational modifications effectively, secretes high levels of heterologous proteins with minimal endogenous protein secretion, facilitating the downstream protein purification. Additionally, there is a growing collection of synthetic biology tools available for this organism[14–21]. *K. phaffii* presents several benefits compared to the model S. cerevisiae system, such as robust endogenous promoters that facilitate the production of heterologous proteins, enabling recombinant products to constitute up to 30% of the total protein output[17]. Furthermore, *K. phaffii* incorporates shorter and less branched mannose residues, which leads to lower levels of hyperglycosylation in the recombinant products compared to *S. cerevisiae*[22].

As a methylotroph, *K. phaffii* has the ability to use methanol as its sole carbon source. Therefore, this cell has native pathways that are strongly inducible under methanol. *K. phaffii* has two alcohol oxidase (AOX) genes (*AOX1* and *AOX2*) that are part of the initial enzymatic steps to assimilate methanol to formaldehyde. The *AOX1* constitutes the majority of AOX activity in K. phaffii since *AOX1*-deficient cells grow slowly in methanol whereas *AOX2* mutants have the same growth rate as wildtype cells[23,24]. One of the most important characteristics of this cell is the very strong and tightly regulated promoter region of its *AOX1* that has been used widely in the literature for controlled recombinant protein production[25–27]. Over the past few decades, three primary types of *K. phaffii* host strains have been utilized, with differences in their ability to metabolize methanol: 1- the wildtype cells containing both copies of *AOX1* and *AOX2*, which grow fast on methanol, therefore methanol utilization plus (Mut^{+}). 2- Strains with a deleted *AOX1* which exhibit a methanol utilization slow phenotype (Mut^{S}), and 3- strains with deletions in both *AOX1* and *AOX2* genes, leading to a methanol utilization minus phenotype (Mut^{-}).

1-2 Advancements in tool development for synthetic biology and metabolic engineering of *Komagataella phaffii*

Drawing from the successful engineering of *S. cerevisiae* as a microbial cell factory, three key prerequisites can be identified for building stable cell factories: (i) comprehensive genome information; (ii) effective and precise genome engineering tools; and (iii) an adequate set of genetic components, such as promoters and integration sites, to ensure stable gene expression. The genome of multiple strains of K. *phaffii* has been sequenced and annotated[28,29]. Current synthetic biology tools for *K. phaffii* include:

1. Gene expression vectors: Exogenous gene integration can be achieved through either genome integration or expression from an episomal plasmid. Both single and multiple copy integration systems has been developed for *K. phaffii* facilitating stable, high-level expression of heterologous genes[30–33]. Furthermore, A collection of episomal vectors containing different autonomously replicating sequences (ARSs) has been created and systematically evaluated for their transformation efficiency, copy number, and reproductive stability[34–37].
2. Identification of promoter and terminator regions: Recognizing promoter and terminator regions with varying strength levels is crucial for developing synthetic biology and metabolic engineering tools that enable recombinant protein production with controlled levels. The *AOX1* and *GAP1* promoters are the most common promoters for either inducible or constitutive expression of recombinant

products. Moreover, a set of different promoter regions has been identified with different strength levels on both glucose and methanol for K. phaffii[38–41].

Terminators have been shown to impact mRNA half-life and regulatory effects on transcription termination in *S. cerevisiae*[42]. However, the influence of the terminator region on expression levels of recombinant products is not as well studied in *K. phaffii*. Ito el al characterized 72 terminator endogenous, heterogeneous, and synthetic terminators and found a 17-fold tunable range from the strongest *PpAOX1*t to the least strong *ScGIC1*t[43].

3- Genome-editing tools: in recent years, CRISPR-Cas9 has been the most widely used genetic-engineering tool in microbial cell factories[44–47]. This technology uses the RNA-guided Cas9 endonuclease to create double-stranded breaks in the genome in a precise manner. This technology uses cells native DNA repair mechanisms including non-homologous end joining (NHEJ), to create random mutations, or homologous recombination (HR), allowing for precise genetic manipulation such as gene insertions or deletions. Various studies have been conducted to optimize the CRISPR-Cas9 system in *K. phaffii.* Weninger et al. optimized the CRISPR/Cas9 system for precise genome editing in *K. phaffii*, testing various Cas9 coding sequences, gRNA structures, and promoters. Out of 95 combinations, only 6 were functional, indicating the need for further optimizations[48]. Dalvie et al developed a sequencing-based approach to design host-specific hybrid RNA polymerase III promoter regions for efficient gRNA expression, achieving genome editing efficiencies of up to 95%[15]. They also applied this technique to create a multiplexed sgRNA expression system capable of expressing up to three separate sgRNAs simultaneously.

4- Fine-tuning gene expression: Adjusting gene expression is an effective approach for optimizing cellular metabolism while maintaining the activity of essential pathways necessary for cellular function. The CRISPR toolbox in yeast has also been expanded to include gene regulation, by mutating the nuclease domains of Cas9. This will lead to deactivation of the Cas9 nuclease activity (dCas), while still taking advantage of Cas9's precise targeting quality. When the deactivated Cas9 is targeted to promoter regions, it can block transcription, a method known as CRISPR interference (CRISPRi) for gene repression. Repression efficiency can be even increased by fusing transcriptional repressors like Mxi1 or KRAB. Similarly, fusing activation domains like VP64 or VPR to the deactivated Cas9 allows gene overexpression through CRISPR activation (CRISPRa). This approach introduces new possibilities for transcriptional regulation in non-model microbes, where promoter characterization is less advanced[49–51]. CRISPRi and CRISPRa systems have been successfully developed for *K. phaffii*[34,52,53].

5- Homologous recombination machinery engineering: efficient and precise genetic engineering is a prerequisite for making stable cell factories. Seamless gene insertion or deletion is derived from homologous recombination. In contrast to the model yeast *S. cerevisiae*, which predominantly uses HR and requires only short homology arms (<50 bp) for precise gene knock-ins, DNA repair in most non-conventional yeasts typically occurs via NHEJ. As a result, gene integrations in these yeasts often require long homology arms (~1 kb), and inactivation of the native NHEJ pathway, usually by disrupting the KU70 and KU80 genes[54–58]. Other approaches to improve HR includes overexpression of HR-related genes including *RAD51*, *RAD52*, and *RAD54*. Cai et al showed that the HR machinery in *K. phaffii* can be improved by up to nearly 90% by overexpression of PpRAD52[38].

Although these tools and strategies collectively pave the way for more efficient and scalable production of recombinant proteins in *K. phaffii*, it still lacks advanced, high-throughput forward genetics tools that are crucial for constructing efficient cell factories. Additionally, *K. phaffii*'s biology is not as well-understood compared to the model organism *S. cerevisiae.* These tools facilitate the enhancement of the design-build-test-learn cycle and enable the rapid identification of novel mutations that influence various phenotypes.

Genome-wide pooled CRISPR screens are conducted by introducing various genetic perturbations into a pool of cells. This is primarily achieved by expressing Cas9 endonuclease in the cells alongside a sgRNA library that targets all genes in the genome, thereby inducing mutations. The mutated cells are then allowed to grow, either under normal conditions or in the presence of a specific biological challenge, which selects for mutations that enhance a desired phenotype, such as tolerance to stress conditions like extreme temperatures, varying salt concentrations, pH levels, or toxic compounds. The surviving mutants are then evaluated by next-generation sequencing (NGS). This will lead to discovering novel phenotypes and identifying essential and non-essential genes. While these techniques have not yet been widely applied to non-model organisms like *K. phaffii*, the growing adoption of CRISPR systems in these organisms holds great potential for advancing their genetic and metabolic understanding, ultimately enhancing their utility in industrial applications.

1-3 Protein Secretion from *Komagataella phaffii*: bottlenecks of the secretory pathway

The primary challenges in recombinant protein secretion involve limitations in the secretory pathway as well as proper folding of the recombinant product[11,59,60]. These factors greatly affect the titers of the product which is desired to be secreted to facilitate the downstream purification steps. The key differences between protein secretion pathways of various yeast species have been compared by Delic et al[61]. In several instances, *K. phaffii* outperforms *S. cerevisiae* in secretion yields of recombinant products, often due to higher biomass accumulation. Similar to what has been observed in *S. cerevisiae*, it is estimated that about 10% of the genes in *K. phaffii's* genome are involved in the secretory pathway, including those associated with (i) the endoplasmic reticulum (ER), (ii) protein folding, (iii) glycosylation, (iv) proteolytic processing, (v) the ER-associated degradation (ERAD) pathway, (vi) the Golgi apparatus, (vii) SNAREs, and (viii) other components of vesicle-mediated transport.

The eukaryotic protein secretion pathway predominantly follows the ER–Golgi route. This is initiated by the translocation of proteins across the ER membrane, facilitated by the addition of a signal peptide at the N-terminus of the newly synthesized polypeptide ensuring its precise localization. The first major bottleneck in this pathway is the translocation of the proteins from the cytoplasm into the ER lumen. This complex process involves a network of proteins that oversee the targeting of the protein to the ER membrane, its translocation, folding, post-translational modifications, quality control, and trafficking. The membrane targeting depends on the hydrophobicity and amino acid composition of the fully translated signal peptide and it occurs either co-translationally, that depends on the recognition of N-terminal hydrophobic signal sequences of the nascent chain protein by a signal recognition particle (SRP) as its being translated by a ribosome[26,62–65], or post-translationally that is independent of SRPs and ribosomes[66].

In yeast, two distinct translocation pores are present: the Sec61 and the Ssh1 complex. The Sec61 translocation complex is comprised of Sec61, Sbh1, and Sss1. Sec61 and Sss1 have conserved sequences and are essential for protein translocation and cell survival, whereas Sbh1 is not critical for protein translocation in *S. cerevisiae*. The translocon pore is most probably formed by dimers or trimers of the Sec61 complex. During post-translational translocation, the Sec61 complex combines with Sec62, Sec63, Sbh1, Sec71, and Sec72 to form the heptameric SEC complex. For co-translational translocation the hexameric SEC′ complex forms, which consists of the Sec61 complex and Sec63, Sec71, and Sec72. Additionally, the heterotrimeric Ssh1 complex consists of Ssh1 (the non-essential homolog of Sec61), Sbh2 and the Sss1 subunits and is mainly involved in co-translational translocation. While Sec61 translocon can process a wide range of signal sequences, Ssh1 translocon only accepts a limited number of signal sequences including hydrophobic signal sequences such as Kar2 and invertase.

Co-translational translocation into the ER lumen in *S. cerevisiae* involves the interaction of the nascent protein with ribosome, signal recognition particle (SRP), signal recognition receptor (SR), and either the Ssh1 or Sec61 translocon pores. SRP is a complex of six proteins and a 7S RNA, with most components assembling in the nucleus before binding the Srp54 subunit in the cytosol. Srp54 recognizes the signal sequence of the ribosome nascent chain complex (RNC), forming the SRP-RNC complex and pausing translation until the ribosome binds to the translocon pore. In post-translational translocation, after release of the nascent chain from the ribosome, polypeptides should remain unfolded to avoid aggregation. This happens by binding to cytosolic chaperones Ssa1 and Ydj1, which are released just before the nascent chain translocate into the ER. Sec62 then recognizes and binds to the signal peptide of nascent proteins.

Translocation is supported by molecular chaperones including Sec63 and Kar2 that also assist in oxidative protein folding. Sec63 stabilizes the post-translational SEC complex by binding Sec62, gates the translocon pore with Kar2, aiding in assembling the SEC and SEC' translocon complexes, and participating in co-translational transport via the Ssh1 pore. Kar2, along with co-chaperones Lhs1 and Sil1, helps pull nascent polypeptides into the ER, with Lhs1 being the primary nucleotide exchange factor. Sil1's role in translocation is less clear, particularly during normal growth, but it promotes Kar2 recruitment and ATPase activation by Sec63.

Proteins that fail to translocate can sequester cytosolic chaperones involved in translocation, such as Ssa1, Ssa2, Ssb1, and Ssb2. Although these chaperones typically aid in translocation under normal conditions, overexpression during bioproduction might create an unstable cellular environment. Improper translocation into the ER leads to misfolded folding. Protein folding is an ATP-dependent process and involves ER-resident proteins like Kar2, Scj1, Pdi1, Ero1, and Jem1. Misfolded proteins trigger the unfolded protein response (UPR) and are subsequently degraded via the ER-associated degradation (ERAD) pathway.

The efficiency of recombinant protein secretion in yeast is significantly influenced by the proper functioning of the secretory pathway and the correct folding of proteins. The process of translocation into the ER, whether co-translational or post-translational, involves requires coordination between translocon pores, chaperones, and other ER-resident proteins. Failure in this process can result in protein misfolding triggering stress responses like UPR, which can impact the efficiency of heterologous protein production.

High-throughput ribosome profiling under heterologous conditions offers a precise method to measure global mRNA translation at any given time. Using this method in antibody-producing CHO cells, Kallehauge et al have shown that the recombinant antibody was the most abundant transcript, occupying up to 15% of translating ribosomes, and improved protein production by knocking down the unnecessary, highly expressed Neo^R gene[67]. Therefore, highly expressed host proteins that are translocated into the ER can hinder the translocation of heterologous proteins due to the limited availability and efficiency of Sec-translocons and protein folding chaperones. Thus, identifying and targeting the genes that are highly expressed and are translocated in ER could be a rational approach for enhancing the production of recombinant proteins in yeast systems.

1-4 Thesis organization

The first chapter offers an overview of how biotechnology plays a role in tackling major global issues like healthcare, climate change, energy and food security. It focuses on the increasing demand for biologics and the advancements made in producing proteins. The chapter emphasizes the benefits of using non-conventional hosts, particularly *Komagataella phaffii*, for the efficient and scalable production of these proteins. While significant progress has been made in developing synthetic biology tools for *K. phaffii*, it points out that our understanding of its biology lags that of *S. cerevisiae* and it still lacks advanced, high-throughput forward genetics tools essential for enhancing the design-build-test-learn cycle and rapidly identifying novel mutations that influence various phenotypes. The chapter also discusses the bottlenecks in the secretory pathway of *K. phaffii*, particularly during the ER translocation and folding of recombinant proteins, and discusses strategies to improve the efficiency of protein secretion. By addressing these obstacles, the chapter highlights how *K. phaffii* holds promise as a platform for industrial biotechnology applications.

The second chapter focuses on the design and application of single guide RNA (sgRNA) libraries for CRISPR-Cas9 genome-wide screening in non-conventional microbial hosts. The chapter provides a detailed protocol, including Python scripts, for creating an sgRNA library that covers all genes in the genome. The chapter also discusses the challenges of sgRNA design in non-model hosts, the importance of optimizing sgRNA targeting efficiency, and the use of computational tools like CHOPCHOP v3 for designing highly active sgRNA libraries. This approach facilitates the development of advanced synthetic biology tools and enhances the potential application of non-conventional hosts in biotechnological applications.

Chapter 3 introduces a new, powerful tool for studying *Komagataella phaffii*: a high-activity CRISPR-Cas9 genome-wide sgRNA library. This tool facilitates functional genomic screening in this non-conventional yeast and expands on the advanced synthetic biology tools available for this yeast. By designing and validating a high-activity CRISPR-Cas9 genome-wide sgRNA library, the chapter uncovers unique essential genes specific to *K. phaffii*, providing deeper insights into its biology and identifying promising targets for optimizing the strain, especially for complex phenotypes such as recombinant protein secretion and glycosylation.

Chapter 4 explores the translational landscape of *Komagataella phaffii* during heterologous protein expression. By employing ribosome profiling (Ribo-seq) and next-generation sequencing, we study protein synthesis in *K. phaffii* before and after methanol induction using an engineered strain that is able to produce and secrete human serum albumin (HSA). As a rational approach to enhance bioproduction, we identify non-essential host cell proteins that consume significant biogenetic resources, particularly in the early secretory pathway. We then apply these findings to rationally engineer *K. phaffii* for increases secretion of HSA and find that a combination of knockouts can improve this complex phenotype.

Finally, in chapter 5 the findings of this dissertation are summarized, broader implications of this dissertation on this specific field are discussed, and potential future direction is suggested.

References

1. *Biologics Market Size Worth $399.5 Billion By 2025 | Growth Rate: 3.9%*. (Grand View Research, Inc).
2. Sanchez-Garcia, L. *et al.* Recombinant pharmaceuticals from microbial cells: a 2015 update. *Microb. Cell Fact.* **15**, 33 (2016).
3. Carlson, R. Estimating the biotech sector's contribution to the US economy. *Nat. Biotechnol.* **34**, 247–255 (2016).
4. Wang, G., Huang, M. & Nielsen, J. Exploring the potential of Saccharomyces cerevisiae for biopharmaceutical protein production. *Curr. Opin. Biotechnol.* **48**, 77–84 (2017).
5. Love, K. R., Dalvie, N. C. & Love, J. C. The yeast stands alone: the future of protein biologic production. *Curr. Opin. Biotechnol.* **53**, 50–58 (2018).
6. 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).
7. Wagner, J. M. & Alper, H. S. Synthetic biology and molecular genetics in non-conventional yeasts: Current tools and future advances. *Fungal Genet. Biol.* **89**, 126–136 (2016).
8. Bustos, C. *et al.* Advances in cell engineering of the Komagataella phaffii platform for recombinant protein production. *Metabolites* **12**, 346 (2022).
9. Mastropietro, G., Aw, R. & Polizzi, K. M. Expression of proteins in Pichia pastoris. in *Methods in Enzymology* 53–80 (Elsevier, 2021).
10. Ergün, B. G., Berrios, J., Binay, B. & Fickers, P. Recombinant protein production in Pichia pastoris: from transcriptionally redesigned strains to bioprocess optimization and metabolic modelling. *FEMS Yeast Res.* **21**, (2021).
11. Karbalaei, M., Rezaee, S. A. & Farsiani, H. Pichia pastoris: A highly successful expression system for optimal synthesis of heterologous proteins. *J. Cell. Physiol.* **235**, 5867–5881 (2020).
12. De Wachter, C., Van Landuyt, L. & Callewaert, N. Engineering of Yeast Glycoprotein Expression. in *Advances in Biochemical Engineering/Biotechnology* 93–135 (Springer International Publishing, Cham, 2018).
13. Corchero, J. L. *et al.* Unconventional microbial systems for the cost-efficient production of high-quality protein therapeutics. *Biotechnol. Adv.* **31**, 140–153 (2013).
14. Tafrishi, A. *et al.* Functional genomic screening in Komagataella phaffii enabled by high-activity CRISPR-Cas9 library. *Metab. Eng.* **85**, 73–83 (2024).

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

16. Weis, R. High-Throughput Screening and Selection of Pichia pastoris Strains. *Methods Mol. Biol.* **1923**, 169–185 (2019).

17. Gao, J., Jiang, L. & Lian, J. Development of synthetic biology tools to engineer Pichia pastoris as a chassis for the production of natural products. *Synth. Syst. Biotechnol.* **6**, 110–119 (2021).

18. Demir, İ. & Çalık, P. Hybrid-architectured double-promoter expression systems enhance and upregulate-deregulated gene expressions in Pichia pastoris in methanol-free media. *Appl. Microbiol. Biotechnol.* **104**, 8381–8397 (2020).

19. García-Ortega, X. *et al.* Rational development of bioprocess engineering strategies for recombinant protein production in Pichia pastoris (Komagataella phaffii) using the methanol-free GAP promoter. Where do we stand? *N. Biotechnol.* **53**, 24–34 (2019).

20. Liu, Q. *et al.* CRISPR-Cas9-mediated genomic multiloci integration in Pichia pastoris. *Microb. Cell Fact.* **18**, 144 (2019).

21. Vogl, T. *et al.* Methanol independent induction in Pichia pastoris by simple derepressed overexpression of single transcription factors. *Biotechnol. Bioeng.* **115**, 1037–1050 (2018).

22. Valli, M. *et al.* A subcellular proteome atlas of the yeast Komagataella phaffii. *FEMS Yeast Res.* **20**, (2020).

23. Zhang, H. *et al.* Alcohol oxidase (AOX1) from Pichia pastoris is a novel inhibitor of prion propagation and a potential ATPase. *Mol. Microbiol.* **71**, 702–716 (2009).

24. de Hoop, M. J. *et al.* Overexpression of alcohol oxidase in Pichia pastoris. *FEBS Lett.* **291**, 299–302 (1991).

25. Gasser, B., Maurer, M., Gach, J., Kunert, R. & Mattanovich, D. Engineering of Pichia pastoris for improved production of antibody fragments. *Biotechnol. Bioeng.* **94**, 353–361 (2006).

26. Crowell, L. E. *et al.* On-demand manufacturing of clinical-quality biopharmaceuticals. *Nat. Biotechnol.* **36**, 988–995 (2018).

27. Dalvie, N. C. *et al.* Engineered SARS-CoV-2 receptor binding domain improves manufacturability in yeast and immunogenicity in mice. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2106845118 (2021).

28. Love, K. R. *et al.* Comparative genomics and transcriptomics of Pichia pastoris. *BMC Genomics* **17**, (2016).

29. Alva, T. R., Riera, M. & Chartron, J. W. Translational landscape and protein biogenesis demands of the early secretory pathway in Komagataella phaffii. *Microb. Cell Fact.* **20**, 19 (2021).

30. Yoshimasu, M. A., Ahn, J.-K., Tanaka, T. & Yada, R. Y. Soluble expression and purification of porcine pepsinogen from Pichia pastoris. *Protein Expr. Purif.* **25**, 229–236 (2002).

31. Ha, S. H. *et al.* Molecular cloning and high-level expression of G2 protein of hantaan (HTN) virus 76-118 strain in the yeast Pichia pastoris KM71. *Virus Genes* **22**, 167–173 (2001).

32. Wang, Y., Wang, J., Leng, F., Ma, J. & Bagadi, A. Expression of Aspergillus niger glucose oxidase in Pichia pastoris and its antimicrobial activity against Agrobacterium and Escherichia coli. *PeerJ* **8**, e9010 (2020).

33. Papakonstantinou, T., Harris, S. & Hearn, M. T. W. Expression of GFP using Pichia pastoris vectors with zeocin or G-418 sulphate as the primary selectable marker. *Yeast* **26**, 311–321 (2009).

34. Yang, Y. *et al.* High efficiency CRISPR/Cas9 genome editing system with an eliminable episomal sgRNA plasmid in Pichia pastoris. *Enzyme Microb. Technol.* **138**, 109556 (2020).

35. Gu, Y. *et al.* Construction of a series of episomal plasmids and their application in the development of an efficient CRISPR/Cas9 system in Pichia pastoris. *World J. Microbiol. Biotechnol.* **35**, 79 (2019).

36. Camattari, A. *et al.* Characterization of a panARS-based episomal vector in the methylotrophic yeast Pichia pastoris for recombinant protein production and synthetic biology applications. *Microb. Cell Fact.* **15**, 139 (2016).

37. Cregg, J. M., Barringer, K. J., Hessler, A. Y. & Madden, K. R. Pichia pastoris as a host system for transformations. *Mol. Cell. Biol.* **5**, 3376–3385 (1985).

38. Cai, P. *et al.* Recombination machinery engineering facilitates metabolic engineering of the industrial yeast Pichia pastoris. *Nucleic Acids Res.* **49**, 7791–7805 (2021).

39. Xu, N. *et al.* Identification and characterization of novel promoters for recombinant protein production in yeast Pichia pastoris. *Yeast* **35**, 379–385 (2018).

40. Yurimoto, H., Oku, M. & Sakai, Y. Yeast methylotrophy: metabolism, gene regulation and peroxisome homeostasis. *Int. J. Microbiol.* **2011**, 101298 (2011).

41. Zhang, A.-L. *et al.* Recent advances on the GAP promoter derived expression system of Pichia pastoris. *Mol. Biol. Rep.* **36**, 1611–1619 (2009).

42. Curran, K. A., Karim, A. S., Gupta, A. & Alper, H. S. Use of expression-enhancing terminators in Saccharomyces cerevisiae to increase mRNA half-life and improve gene expression control for metabolic engineering applications. *Metab. Eng.* **19**, 88–97 (2013).

43. Ito, Y. *et al.* Exchange of endogenous and heterogeneous yeast terminators in Pichia pastoris to tune mRNA stability and gene expression. *Nucleic Acids Res.* **48**, 13000–13012 (2020).

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

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

46. 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).

47. 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).

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

49. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).

50. 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).

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

52. Qiao, S., Bai, F., Cai, P., Zhou, Y. J. & Yao, L. An improved CRISPRi system in Pichia pastoris. *Synth. Syst. Biotechnol.* **8**, 479–485 (2023).

53. Baumschabl, M., Prielhofer, R., Mattanovich, D. & Steiger, M. G. Fine-Tuning of Transcription in Pichia pastoris Using dCas9 and RNA Scaffolds. *ACS Synth. Biol.* **9**, 3202–3209 (2020).

54. Weninger, A. *et al.* Expanding the CRISPR/Cas9 toolkit for Pichia pastoris with efficient donor integration and alternative resistance markers. *J. Cell. Biochem.* **119**, 3183–3198 (2018).

55. Näätsaari, L. *et al.* Deletion of the Pichia pastoris KU70 homologue facilitates platform strain generation for gene expression and synthetic biology. *PLoS One* **7**, e39720 (2012).

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

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

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

59. Bernauer, L., Radkohl, A., Lehmayer, L. G. K. & Emmerstorfer-Augustin, A. Komagataella phaffii as Emerging Model Organism in Fundamental Research. *Front. Microbiol.* **11**, 607028 (2020).

60. Puxbaum, V., Mattanovich, D. & Gasser, B. Quo vadis? The challenges of recombinant protein folding and secretion in Pichia pastoris. *Appl. Microbiol. Biotechnol.* **99**, 2925–2938 (2015).

61. Delic, M. *et al.* The secretory pathway: exploring yeast diversity. *FEMS Microbiol. Rev.* **37**, 872–914 (2013).

62. Love, K. R. *et al.* Systematic single-cell analysis of Pichia pastoris reveals secretory capacity limits productivity. *PLoS One* **7**, e37915 (2012).

63. Zahrl, R. J., Mattanovich, D. & Gasser, B. The impact of ERAD on recombinant protein secretion in Pichia pastoris (syn Komagataella spp.). *Microbiology* **164**, 453–463 (2018).

64. Chartron, J. W., Hunt, K. C. L. & Frydman, J. Cotranslational signal-independent SRP preloading during membrane targeting. *Nature* **536**, 224–228 (2016).

65. Nyathi, Y., Wilkinson, B. M. & Pool, M. R. Co-translational targeting and translocation of proteins to the endoplasmic reticulum. *Biochim. Biophys. Acta* **1833**, 2392–2402 (2013).

66. Zimmermann, R., Eyrisch, S., Ahmad, M. & Helms, V. Protein translocation across the ER membrane. *Biochim. Biophys. Acta Biomembr.* **1808**, 912–924 (2011).

67. Kallehauge, T. B. *et al.* Ribosome profiling-guided depletion of an mRNA increases cell growth rate and protein secretion. *Sci. Rep.* **7**, 40388 (2017).

Chapter 2: sgRNA library design for CRISPR-Cas9 genome-wide screening in non-conventional hosts

2.1 Abstract

High-throughput forward genetic screens, powered by sequencing advancements, are essential for understanding cell biology and identifying targets for strain engineering. Recent advancements in genetic engineering have enabled systematic investigations into gene function and genotype-to-phenotype association. This is particularly important when working with non-conventional hosts, as these non-model organisms often exhibit desirable and industrially relevant phenotypes. Nonetheless, to harness these hosts' potential, novel synthetic biology tools are needed for precise genome engineering to create advanced microbial cell factories.

The application of CRISPR-Cas9 technology, in combination with pooled single guide RNA (sgRNA) libraries, has revolutionized the ability to conduct such analyses across a broad range of microbial hosts. This approach allows for the systematic disruption of genes and the subsequent analysis of resulting phenotypes, providing valuable insights into gene function. Key considerations in this process include the optimization of sgRNA design to ensure maximum efficiency, minimizing off-target effects, and ensuring unique targeting within the genome. In this methods chapter, we present a detailed protocol for designing sgRNA libraries for a genome-wide CRISPR-Cas9 screen, targeting specific regions of every gene in the genome. This protocol offers an efficient framework for creating sgRNA libraries for genome-wide CRISPR-Cas9 screens.

2.2 Introduction

The advent of the robust, high-fidelity, and programmable CRISPR-Cas9 system has significantly advanced genetic manipulation. By introducing a targeted double-stranded break in the genome via the Cas9 endonuclease and leveraging the cell's DNA repair mechanisms, it enables precise gene disruptions, including creating mutations or gene insertions and deletions. Genome-wide single guide RNA (sgRNA) library screens, utilizing a functional Cas9 protein alongside a pooled collection of sgRNAs, are widely used for various purposes including functional genomics, the discovery of novel phenotypes, and the identification of essential and non-essential genes[1–5]. Performing genome-wide loss-of-function screens has now been facilitated by the advancement of synthetic biology and DNA sequencing and synthesis. This approach involves applying selective pressure on cells to isolate those that exhibit a desired phenotype. Cells surviving in the pressure condition can then be studied separately to uncover the genetic basis of the observed phenotype[6]. In this case, a pooled library of sgRNAs is transformed into cells expressing Cas9 under a selective pressure inducing mutations across the genome leading to disrupting gene functions. These mutations would cause either an enrichment or a depletion of some of the sgRNAs targeting specific genes. Investigating these genes' functions can be a promising approach in associating genotypes to phenotypes[7,8].

Although many non-model hosts address bioproduction technical challenges including beneficial native phenotypes, environmental stress tolerance, and an inherent ability to produce desired products, they are often overlooked as microbial hosts due to lack of existing advanced synthetic biology tools[9]. Over the past few years, several sgRNA targeting-efficiency predictor tools have been developed, primarily based on model organisms such as mammalian cells, *Saccharomyces cerevisiae*, and *Escherichia coli*[10–15]. Since these tools are trained on data from model hosts, their accuracy in predicting active guides for non-model hosts remains uncertain[16,17]. To address this limitation, one approach is to target each gene with multiple sgRNAs, potentially with various targeting efficiencies. This redundancy increases the likelihood of including at least one active guide, thereby improving the overall effectiveness of the screen.

We have previously developed detailed experimental protocols to create activity profiles for every sgRNA in a library. These profiles can be used in conjunction with functional screens to enhance screening outcomes. The protocol involves deactivating the dominant DNA repair pathway (non-homologous end joining (NHEJ) for non-conventional yeasts) and performing growth screens in the presence of an active Cas9. In the absence of NHEJ and without a DNA repair template, active sgRNAs induce cell death, leading to their depletion from the sgRNA population. Hence, the abundance of sgRNAs in this background can be utilized as a quantitative measure of guide activity[1,2,18].

In this protocol chapter, we outline a method and provide Python scripts for designing an n-fold coverage library targeting every feature (e.g., gene, coding sequence, exon, etc.) of a species. This approach is particularly useful for non-conventional hosts that lack appropriate sgRNA design tools. Using this method, we have successfully created libraries for *Komagataella phaffii* and *Kluyveromyces marxianus* and have validated the design process for *Komagataella phaffii*[2]. This protocol offers a versatile solution for researchers working with diverse microbial hosts, facilitating the generation of comprehensive sgRNA libraries and development of advanced synthetic biology tools for non-conventional hosts.

2.3 Materials

2.3.1 Software and computer

python 2.7.14

scipy 1.2.2

gffutils 0.10.1

mySQL-python 1.2.3

numpy 1.16.6

pandas 0.24.2

scikit-learn 0.18.1

scipy 1.2.2

Laptop or desktop computer that meets the requirements to run python 2.7.14

2.4 Methods

2.4.1 Experimental pipeline of pooled CRISPR-Cas9 genome-wide screening in non-conventional hosts

Fig. 2-1 presents a schematic overview of the experimental pipeline for pooled CRISPR-Cas9 knockout screens. After selecting a host organism, an n-fold coverage sgRNA library is designed to target relevant protein-coding genes, with the option to include non-coding regions such as promoters and introns. The library is synthesized as an oligonucleotide pool, cloned into a plasmid backbone, and transformed into cells under selective pressure. Cells are cultured in the selective media for several days, with sub-culturing into fresh media to ensure proper library distribution. Plasmids are then extracted and quantified via qPCR to confirm sufficient material for 100-fold library coverage. The sgRNA regions along with NGS adapters are then PCR-amplified and sequenced. Sequencing data is used to calculate sgRNA abundance, from which the importance of each targeted gene to cell health is determined.

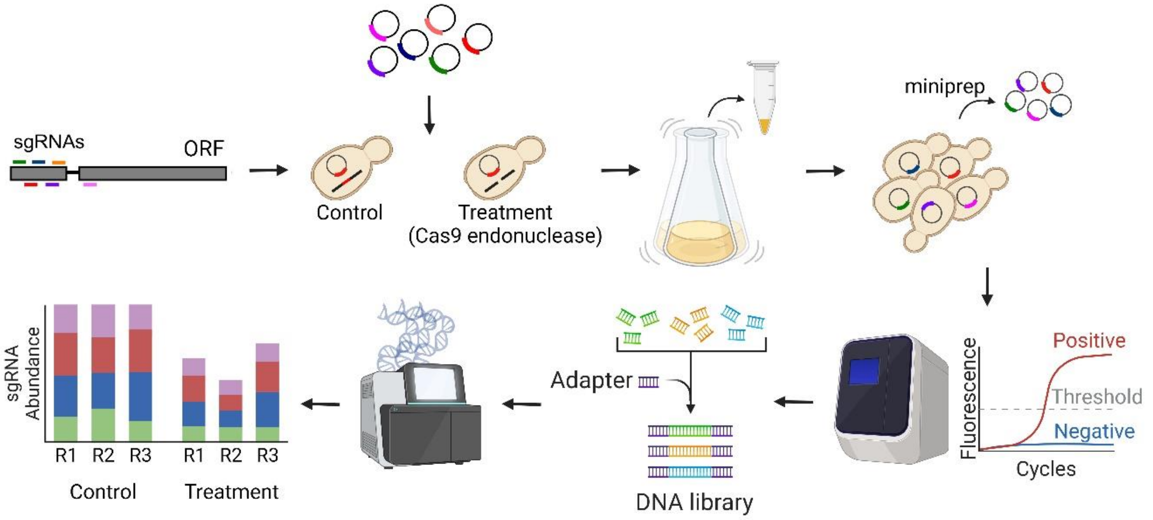


Fig. 2- 1. Experimental workflow for conducting pooled CRISPR-Cas9 knockout screens. A library of sgRNAs targeting every open reading frame (ORF) in the genome is designed, synthesized, and cloned into a plasmid backbone, which is then transformed into both the control and the treatment strains of interest. The cells are cultured under selection pressure, causing perturbations in sgRNA abundance. These changes can be measured to assess the significance of the genes targeted by the sgRNAs. After the screening process, plasmids are extracted from the cells, quantified by qPCR, and sequenced using NGS to draw conclusions about gene importance based on the observed perturbations. The figure is created with BioRender.com.

2.4.2 CRISPR vectors for CRISPR-Cas9 gene editing

A plasmid with the following features is required for conducting pooled CRISPR-Cas9 genome-wide screens:

1. Cas9 expression: A functional Cas9 protein driven by a constitutive promoter.
2. sgRNA expression cassette: An efficient sgRNA expression system. To produce a 20 bp mature sgRNA it is suggested to use a hybrid promoter including an RNA Pol-III promoter and tRNA sequences to leverage the cell's endogenous tRNA processing. The sgRNA maturation process is crucial in efficient Cas9 targeting[19–21].
3. Selection marker: Markers for plasmid maintenance and selection during library amplification in *E. coli* and transformation in the non-conventional host.
4. Origin of replication: An origin of replication that ensures plasmid stability in the host cell is crucial. In our library validation workflow, we evaluate sgRNA abundance in strains with integrated Cas9, either in a wildtype background—where a fitness score (FS) is calculated to measure the impact of CRISPR-induced double-stranded breaks on cell viability—or in a background where the main DNA repair mechanism is deactivated, where a cutting score (CS) is calculated to assess sgRNA activity. It is essential to ensure plasmid stability so that any observed plasmid loss is attributed exclusively to Cas9 activity, rather than plasmid instability. Additionally, a separate origin of replication is required for plasmid transformation in *E. coli.*

2.4.3 CRISPR-Cas9 genome-wide sgRNA library design

The following protocol provides a comprehensive explanation as well as Python scripts for designing an n-fold coverage sgRNA library for CRISPR-Cas9 screens, capable of targeting specific regions within each feature of the genome (i.e., gene, coding sequence, exon, etc.) of any organism of interest. This process requires genome sequences (FASTA format) and genome annotation features (GFF3 format) of the organism (**Fig. 2-2**). We have previously demonstrated the effectiveness of this library design method for non-conventional hosts, by designing and experimentally validating the activity of a 6-fold coverage sgRNA library, targeting the first 300 bp of each coding sequence, designed for *Komagataella phaffii* GS115 strain. The designed library contained a total of 31,634 sgRNAs. Experimental validation showed that 98.7% of the sgRNAs were active, with 75.6% classified as highly active[2].

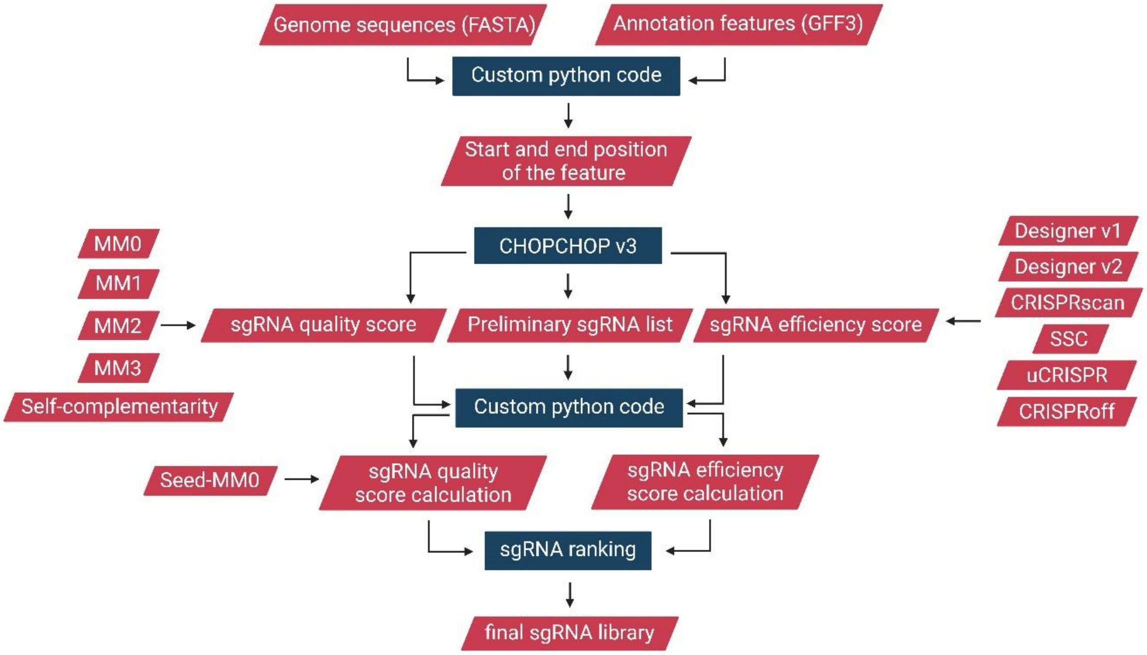


Fig. 2- 2. Schematic flowchart of the design process of the n-fold coverage sgRNA library for genome-wide CRISPR-Cas9 screens. Custom Python scripts are used to determine the target locations for each feature (i.e., gene, coding sequence, exon, etc.), which are then submitted to CHOPCHOP v3[22]. This tool identifies all sgRNAs within the specified regions and provides efficiency predictions from various methods, along with information on the uniqueness of each sgRNA. Additional Python scripts interpret this efficiency and quality data to generate a single quality and efficiency score for each sgRNA. Based on these scores, sgRNAs for each gene are ranked, and the top n sgRNAs are selected for inclusion in the final library.

2.4.3.1 Obtaining the genome sequence and annotation features files

To begin, download the genomic data for the organism of interest from NCBI. Follow these steps:

1- Navigate to the NCBI website and select "Genome" from the dropdown menu on the left side of the search box.

2- In the search box, enter the scientific name of the organism or the specific strain you are interested in and click "Search".

3- On the resulting page, identify and open the appropriate "Assembly" for the strain that matches your criteria, including the scientific name and relevant modifiers.

4- In the “Download” options, select "Genome sequences (FASTA)" and "Annotation features (GFF)."

5- Finally, click "Download" to obtain the genomic data files.

2.4.3.2 Determining the target location and coverage

To enhance the likelihood of creating a library biased towards more active sgRNAs, multiple sgRNAs are often designed to target the same region. While this approach improves the chances of obtaining at least one active guide per feature, it results in a significantly larger library. This expansion introduces complexities in both the experimental aspects, especially if the organism of interest lacks efficient transformation protocols, and the downstream data analysis processes. Therefore, it is crucial to identify a highly efficient transformation protocol for the organism of interest before moving on to the library design process.

To accurately target sgRNAs to specific regions in the genome, we first need to determine the exact location of each region. This information is embedded into the genome annotation features file with GFF3 format which is a nine-column, tab-delimited, plain text file. The Python package *gffutils* creates a database from the annotation data, making the information in the GFF3 file easily accessible and facilitating efficient identification of target regions. The code below demonstrates how to generate the database from the GFF3 file, making it easier to access all the relevant information.

```
with open("file_name.gff3", "rt") as file:
      data = file.read()
gffutils.create_db(data, dbfn='microorganism_name.db', force=True,
from_string=True,merge_strategy="merge", sort_attribute_values=True)
db = gffutils.FeatureDB('microorganism_name.db', keep_order=True)
```

To determine the start and end locations of the targeting regions in the genome, the database entries are iterated through, and the specific feature of interest targeted with the library, such as a gene, coding sequence (CDS), or exon, is identified. The feature's chromosome, start, end, and strand information are then saved in a dictionary to further improve accessibility and facilitate iterability. The following script generates the dictionary containing all the data for a feature of type "gene".

```
db_dict = {}
for i in list(db.features_of_type("gene")):
        db_dict[i['ID'][0]] = []
        feature_info = [i.seqid, i.start, i.end, i.strand]
        db_dict[i['ID'][0]].append(feature_info)
```

2.4.3.3 Comprehensive sgRNA library design using CHOPCHOP v3

To identify all possible sgRNAs targeting a specific feature, we used CHOPCHOP v3[22]. The web-based version of CHOPCHOP does not support designing libraries that require reading the entire genome annotation features. For this purpose, it is essential to use the command-line version. The command-line version of CHOPCHOP v3 utilizes both the genome sequence (FASTA) and genome annotation features (GFF3) files to generate a list of sgRNAs, along with scores for uniqueness, self-complementarity, and efficiency predictions. All the necessary requirements and step-by-step installation guides for CHOPCHOP v3 are explained by its authors[22,23].

CHOPCHOP essentially requires three inputs: the chromosome name, along with the start and end positions where all potential sgRNAs are to be identified. The command to run CHOPCHOP has been integrated into custom Python scripts, enabling the iteration through the entire GFF3 file to generate the necessary inputs for CHOPCHOP. Additionally, access to the output of CHOPCHOP is facilitated by reading the CHOPCHOP output as a pandas DataFrame.

CHOPCHOP assigns multiple targeting efficiency predictive parameters to each CRISPR-Cas9 sgRNA derived from various tools including Designer v1[11], Designer v2[12], CRISPRscan[13], SSC[14], and uCRISPR[15], and CRISPRoff[24]. To account for all the targeting efficiency parameters, a cumulative score named 'naïve score' is calculated as the sum of all the normalized targeting efficiency predictions from all methods. This method assigns a unified efficiency score to each sgRNA, considering all the built-in efficiency prediction methods.

CHOPCHOP also analyzes the uniqueness of each 20 bp sgRNA in the genome and creates MM0, MM1, MM2, and MM3 scores determining the number of off-target transcripts for each sgRNA with 0, 1, 2, and 3 mismatches, respectively. As an additional measure, CHOPCHOP calculates self-complementarity scores, predicting the likelihood of the sgRNA forming secondary structures with itself, which potentially leads to reduced targeting efficiency[22,25,26]. Another measure of uniqueness, Seed_MM0, was added by the custom Python scripts which is defined as the number of off-targets with zero mismatches in the seed region of the guide (the last 12 bp of the sgRNA immediately followed by NGG PAM motif). Numerous studies have shown that uniqueness of the seed region is crucial in decreasing the likelihood of off-target effects of Cas9[22,27–30]. To calculate Seed_MM0, bowtie[31] is integrated in the Python scripts to align the seed region to the genome and check the uniqueness of the seed region. A unified 'quality' score is assigned to each sgRNA calculated based on criteria mentioned in **Table 2-1**. A quality score of one signifies the utmost uniqueness in the genome, characterized by a 20 bp sgRNA sequence that is unique with up to three mismatches, has a unique seed sequence in the genome, and is least likely to form a secondary structure with itself. As quality scores increase, the distinctiveness of the 20 bp sgRNA decreases.

To select sgRNAs for inclusion in the final n-fold coverage library, the sgRNAs designed for each feature are first ranked in ascending order according to their quality scores. They are then ranked in descending order based on their naïve scores. The top n sgRNAs are subsequently selected and added to the final library list.

As an example, the following custom Python scripts design sgRNAs for the first 40% of each feature of type 'gene' and choose the best ten sgRNAs to be included in the final library based on their quality and naïve scores.

Table 2- 1. Comprehensive breakdown of quality scores.

Quality score	Criteria
1	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 0, MM3 = 0
2	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 0, MM3 = 1
3	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 0, MM3 = 2
4	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 0, MM3 = 3
5	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 1, MM3 = 0
6	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 1, MM3 = 1
7	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 1, MM3 = 2
8	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 1, MM3 = 3
9	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 2, MM3 = 0
10	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 2, MM3 = 1
11	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 2, MM3 = 2
12	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 2, MM3 = 3
13	Seed_MM0 = 1, MM0 = 1

```
PAM = ['AGG', 'TGG', 'CGG', 'GGG']
total_chopchop = pd.DataFrame()
n_fold_library = pd.DataFrame()

#Running CHOPCHOP v3 to find all the sgRNAs in the first 40% bp of each gene:
for key, value in db_dict.items():
        feature_length = value[2] - value[1]
        if value[3] == '+':
                start_loci = value[1]
                end_loci = value[1] + 0.4 * feature_length
        if value[3] == '-':
                start_loci = value[2] - 0.4 * feature_length
                end_loci = value[2]
        ALL = check_output(["./chopchop.py", "-T", "1", "-M", "NGG", "–
        maxMismatches", "3", "-g", "20", "-G",  "path to .mt files in
        config_local.json" , "-o", "Results" , "-Target", "%s:%d-%d"%(value[0],
        start_loci, end_loci), "--scoringMethod", "ALL"], universal_newlines=True)

        if len(ALL) != 0:
                ALL = ALL.decode("utf-8")
                data = io.StringIO(ALL)
                df = pd.read_csv(data, sep='\t')
                df['Target sequence'] = df['Target sequence'].apply(lambda x :
                x[0:20])

                df['feature ID'] = [str(key)] * df.shape[0]
                df['naive score'] = df['XU_2015'] + df['DOENCH_2014'] +
                df['MORENO_MATEOS_2015'] + 0.01 * df['DOENCH_2016'] + 0.01 *
                df['ALKAN_2018'] + 0.01 * df['ZHANG_2019']

#Calculating Seed_MM0 for each sgRNA using Bowtie:
                df['Seed_sequence'] = df['Target sequence'].apply(lambda x : x[8:20])
                for a, b in df.iterrows():
                        Seed_MM0 = 0
                        for i in PAM:
                                data = check_output(["bowtie", "-a", "-p", "4", "-v",
                                "0", "Path to Bowtie indexed files", "-c",
                                "%s"%(b['Seed_sequence']+i)],
                                universal_newlines=True)

                                if len(data) != 0:
                                        data = data.decode("utf-8")
                                        data = io.StringIO(data)
                                        dataframe = pd.read_csv(data, sep='\t',
                                        header=None, usecols=[0, 1, 2, 3, 4, 5, 6,
                                        7], engine='python')

                                        Seed_MM0 += dataframe.shape[0]
                        df.at[a, "Seed_MM0"] = int(Seed_MM0) – 1
```

```
#Defining the Quality score for each sgRNA:
                for a, b in df.iterrows():
                        if b['Seed_MM0'] == 0 and b['MM0'] == 0 and b['MM1'] == 0 and
                        b['MM2'] == 0 and b['MM3'] == 0 and b['Self-complementarity']
                        == 0:

                                df.at[a, "Quality_score"] = 1
                        elif b['Seed_MM0'] == 0 and b['MM0'] == 0 and b['MM1'] == 0
                        and b['MM2'] == 0 and b['MM3'] == 1 and b['Self-
                        complementarity'] == 0:

                                df.at[a, "Quality_score"] = 2
                        elif b['Seed_MM0'] == 0 and b['MM0'] == 0 and b['MM1'] == 0
                        and b['MM2'] == 0 and b['MM3'] == 2 and b['Self-
                        complementarity'] == 0:

                                df.at[a, "Quality_score"] = 3
                        elif b['Seed_MM0'] == 0 and b['MM0'] == 0 and b['MM1'] == 0
                        and b['MM2'] == 0 and b['MM3'] == 3 and b['Self-
                        complementarity'] == 0:

                                df.at[a, "Quality_score"] = 4
                        elif b['Seed_MM0'] == 0 and b['MM0'] == 0 and b['MM1'] == 0
                        and b['MM2'] == 1 and b['MM3'] == 0 and b['Self-
                        complementarity'] == 0:

                                df.at[a, "Quality_score"] = 5
                        elif b['Seed_MM0'] == 0 and b['MM0'] == 0 and b['MM1'] == 0
                        and b['MM2'] == 1 and b['MM3'] == 1 and b['Self-
                        complementarity'] == 0:

                                df.at[a, "Quality_score"] = 6
                        elif b['Seed_MM0'] == 0 and b['MM0'] == 0 and b['MM1'] == 0
                        and b['MM2'] == 1 and b['MM3'] == 2 and b['Self-
                        complementarity'] == 0:

                                df.at[a, "Quality_score"] = 7
                        elif b['Seed_MM0'] == 0 and b['MM0'] == 0 and b['MM1'] == 0
                        and b['MM2'] == 1 and b['MM3'] == 3 and b['Self-
                        complementarity'] == 0:

                                df.at[a, "Quality_score"] = 8

                        elif b['Seed_MM0'] == 0 and b['MM0'] == 0 and b['MM1'] == 0
                        and b['MM2'] == 2 and b['MM3'] == 0 and b['Self-
                        complementarity'] == 0:

                                df.at[a, "Quality_score"] = 9

                        elif b['Seed_MM0'] == 0 and b['MM0'] == 0 and b['MM1'] == 0
                        and b['MM2'] == 2 and b['MM3'] == 1 and b['Self-
                        complementarity'] == 0:
                                df.at[a, "Quality_score"] = 10

                        elif b['Seed_MM0'] == 0 and b['MM0'] == 0 and b['MM1'] == 0
                        and b['MM2'] == 2 and b['MM3'] == 2 and b['Self-
                        complementarity'] == 0:

                                df.at[a, "Quality_score"] = 11

                        elif b['Seed_MM0'] == 0 and b['MM0'] == 0 and b['MM1'] == 0
                        and b['MM2'] == 2 and b['MM3'] == 3 and b['Self-
                        complementarity'] == 0:

                                df.at[a, "Quality_score"] = 12

                        elif b['Seed_MM0'] == 1 and b['MM0'] == 1:

                                df.at[a, "Quality_score"] = 13


#Ranking sgRNAs designed for each CDS based on their Quality_score and naive_score
                df.sort_values(by=['Quality_score', 'naive_score'], ascending=[True,
                False], inplace=True)

                total_chopchop = pd.concat([total_chopchop, df])

                chopchop_df.to_csv('total_chopchop.csv')

#Choosing the first best ten sgRNAs for the final library
                n_fold_library = pd.concat([n_fold_library, df.head(10)])

                n_fold_library.to_csv('n_fold_library.csv')
```

2.5 Conclusion

The integration of high-throughput forward genetic screens, particularly through the application of CRISPR-Cas9 technology and pooled sgRNA libraries, has revolutionized our ability to explore and manipulate gene functions in both model and non-model organisms. This approach is especially valuable for non-conventional hosts like which exhibit unique industrially relevant phenotypes but have been historically underutilized due to a lack of advanced synthetic biology tools. By developing and validating a detailed protocol for designing and implementing sgRNA libraries, we have provided a robust framework for genome-wide CRISPR-Cas9 screens, enabling precise gene disruptions and the identification of essential genes. This method not only enhances our understanding of the genetic basis of desirable traits in non-model organisms but also opens new avenues for optimizing microbial cell factories for biotechnological applications. Through these advancements, the potential of non-conventional hosts can be fully harnessed, driving innovation in metabolic engineering and strain development.

References

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

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

3. Dong, C. *et al.* A genome-wide CRISPR-Cas9 knockout screen identifies essential and growth-restricting genes in human trophoblast stem cells. *Nat. Commun.* **13**, 2548 (2022).

4. 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).

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

6. Warner, J. R., Reeder, P. J., Karimpour-Fard, A., Woodruff, L. B. A. & Gill, R. T. Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. *Nat. Biotechnol.* **28**, 856–862 (2010).

7. Hart, T. *et al.* High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* **163**, 1515–1526 (2015).

8. Sidik, S. M. *et al.* A genome-wide CRISPR screen in Toxoplasma identifies essential Apicomplexan genes. *Cell* **166**, 1423-1435.e12 (2016).

9. 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).

10. Zhang, G., Luo, Y., Dai, X. & Dai, Z. Benchmarking deep learning methods for predicting CRISPR/Cas9 sgRNA on- and off-target activities. *Brief. Bioinform.* **24**, (2023).

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

12. 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).

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

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

15. 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).

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

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

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

20. 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).

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

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

23. Bitbucket. https://bitbucket.org/valenlab/chopchop/src/master/.

24. 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).

25. Gilpatrick, T. *et al.* Targeted Nanopore Sequencing with Cas9 for studies of methylation, structural variants, and mutations. *bioRxiv* (2019) doi:10.1101/604173.

26. Thyme, S. B., Akhmetova, L., Montague, T. G., Valen, E. & Schier, A. F. Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nat. Commun.* **7**, 11750 (2016).

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

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

29. 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).

30. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821 (2012).

31. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

Chapter 3: Functional genomic screening in *Komagataella phaffii* enabled by high-activity CRISPR-Cas9 library

3.1 Abstract

CRISPR-based high-throughput genome-wide loss-of-function screens are a valuable approach to functional genetics and strain engineering. The yeast *Komagataella phaffii* is a host of particular interest in the biopharmaceutical industry and as a metabolic engineering host for proteins and metabolites. Here, we design and validate a highly active 6-fold coverage genome-wide sgRNA library for this biotechnologically important yeast containing 30,848 active sgRNAs targeting over 99% of its coding sequences. Conducting fitness screens in the absence of functional non-homologous end joining (NHEJ), the dominant DNA repair mechanism in *K. phaffii*, provides a quantitative means to assess the activity of each sgRNA in the library. This approach allows for the experimental validation of each guide's targeting activity, leading to more precise screening outcomes. We used this approach to conduct growth screens with glucose as the sole carbon source and identify essential genes. Comparative analysis of the called gene sets identified a core set of *K. phaffii* essential genes, many of which relate to metabolic engineering targets, including protein production, secretion, and glycosylation. The high activity, genome-wide CRISPR library developed here enables functional

This chapter previously appeared in Metabolic Engineering. The original citation is as follows: Tafrishi, A., Trivedi, V., Xing, Z., Li, M., Mewalal, R., Cutler, S., R., Blaby, I., & Wheeldon, I. (2024). Functional genomic screening in *Komagataella phaffii* enabled by high-activity CRISPR-Cas9 library. Metabolic Engineering, 85, 73–83 .

genomic screening in *K. phaffii*, applied here to gene essentiality classification, and promises to enable other genetic screens.

3.2 Introduction

The methylotrophic yeast *Komagataella phaffii*, formerly known as *Pichia pastoris*, is commonly referred to as the "biotech yeast" because of its widespread adoption within the pharmaceutical and biotechnology industry[1–4]. This microorganism has emerged as an important recombinant protein production host because it is able to grow to high cell densities as it favors respiratory growth compared to fermentative yeasts, secretes significant levels of heterologous protein in the media saving time and cost for downstream purification processes, has a strong alcohol oxidase I (*AOX1*) promoter facilitating controlled expression of recombinant genes, is able to perform post-translational modifications similar to higher eukaryotes, can assimilate a variety of carbon sources including methanol, and is a generally faster, easier and cost-efficient expression host compared to mammalian cell lines[5,6].

Constructing advanced microbial cell factories requires the development of efficient genetic engineering tools. High-throughput, trackable, forward genetic engineering tools accelerate the design-build-test-learn cycle and facilitate the rapid identification of novel mutations responsible for various phenotypes. Previous efforts in *K. phaffii* engineering include the development of integrative gene expression systems through the use of homologous recombination (HR)[7,8] the design of episomal gene expression vectors[9], and the standardization of variable strength promoters[7]. CRISPR-Cas9 has become the preferred engineering method allowing for precise, targeted, and relatively rapid genetic modifications[10–12]. Using CRISPR-Cas9 with pooled single guide RNA (sgRNA) libraries, allowing for genome-wide screens, has been used as a high-throughput method to analyze gene functions, assign genotypes to phenotypes, and identify essential genes[13–17]. Other functional genomic tools include random chemical or transposon mutagenesis. While these methods have been used successfully in various applications[18–22], they can be limited by the random nature of the resulting mutants, which is biased towards longer genes[23]. CRISPR approaches use targeted mutagenesis, ultimately producing a more diverse mutant pool and more accurate screening outcomes[15,24,25].

One of the challenges with genome-wide CRISPR screens, particularly in non-conventional species, is accurate guide activity predictions. While a number of CRISPR guide activity predictors have been developed[26], they are most often trained on a selected number of model species (e.g., *Escherichia coli*, *Saccharomyces cerevisiae* or mammalian cell lines) and the ability to predict active guides in other species is not well established[27,28]. A solution to this problem is to design multiple guides to target a single gene, thus biasing the library toward at least one active guide per gene. This redundancy in guide design, however, introduces complexities in downstream analysis and dramatically increases library size, which can be problematic if efficient transformation protocols are not available for the host of interest. We have addressed this problem by developing an experimental approach to generate genome-wide CRISPR activity profiles that can be used in combination with functional screens to improve screening outcomes[15,25]. The basic principle is to deactivate the native dominant DNA repair mechanism, typically non-homologous end joining (NHEJ) in non-conventional yeasts such as *K. phaffii*[2], and conduct growth screens in the absence of DNA repair. Such screens provide an indirect measure of guide activity as any double stranded break in the genome leads to cell death or a dramatic reduction in cell fitness. The guide activity profiles can be incorporated into the screening analysis pipeline by analytically removing inactive or poorly active guides, thus improving screen accuracy[15].

Here, we design, validate, and deploy a 6-fold coverage, high-activity pooled CRISPR-Cas9 sgRNA library targeting over 99% of the protein-coding sequences in *Komagataella phaffii* GS115. By disabling NHEJ via functional disruption of *KU70*, we first quantify the activity of the library. This guide activity data is used to correct the outcomes of fitness screens and accurately identify essential genes with glucose as the sole carbon source. Analysis of the essential genes revealed a set of essential genes common across a collection of industrially relevant biochemical production hosts and model yeasts, and others that are unique to *K. phaffii*. Identification of essential genes contributes to the overall understanding of *K. phaffii* genetics and enhances gene annotation that will help metabolic engineers create optimized *K. phaffii* production strains. The CRISPR screens used to generate this new data opens new functional genetic screening capabilities for the biotech yeast and promises to enable rapid metabolic engineering workflows.

3.3 Results and Discussion

3.3.1 Pooled sgRNA library enables functional genetic screening in *K. phaffii*

Pooled sgRNA libraries enable forward genetic screens. When transformed into a Cas9 expressing strain, each cell expresses a single sgRNA targeting a gene disruption; outgrowth of the transformants creates a pool of mutant cells with varying phenotypes. The fitness effects due to each sgRNA are quantified by determining a fitness score (FS), the log2 ratio of the normalized abundance of the sgRNA in sample to that of a control strain (**Fig. 3-1a**). Similarly, a cutting score (CS) can be determined for each sgRNA by comparing the normalized abundance of guides in a NHEJ deficient strain to a control strain absent of Cas9. Since no DNA repair template is provided and the cells lack NHEJ, a double-stranded break in the genome results in cell death or a dramatic reduction in cell fitness, thus allowing us to quantify Cas9 activity for a given sgRNA. FS and CS profiles for *K. phaffii* GS115 over a six-day period, including one subculture at day 3, are shown in **Fig. 3-1b** and Supplementary Data 1. Fitness effects are evident after three days of growth (the first time the cultures reached confluency) and are more pronounced after subculturing the population and allowing for additional outgrowth. Notably, non-targeting controls consistently exhibit low CS and high FS values across both time points, indicating their inactivity and negligible impact on cell fitness. In contrast, targeting sgRNAs exhibit a range of CS values and have a broad effect on cell fitness.

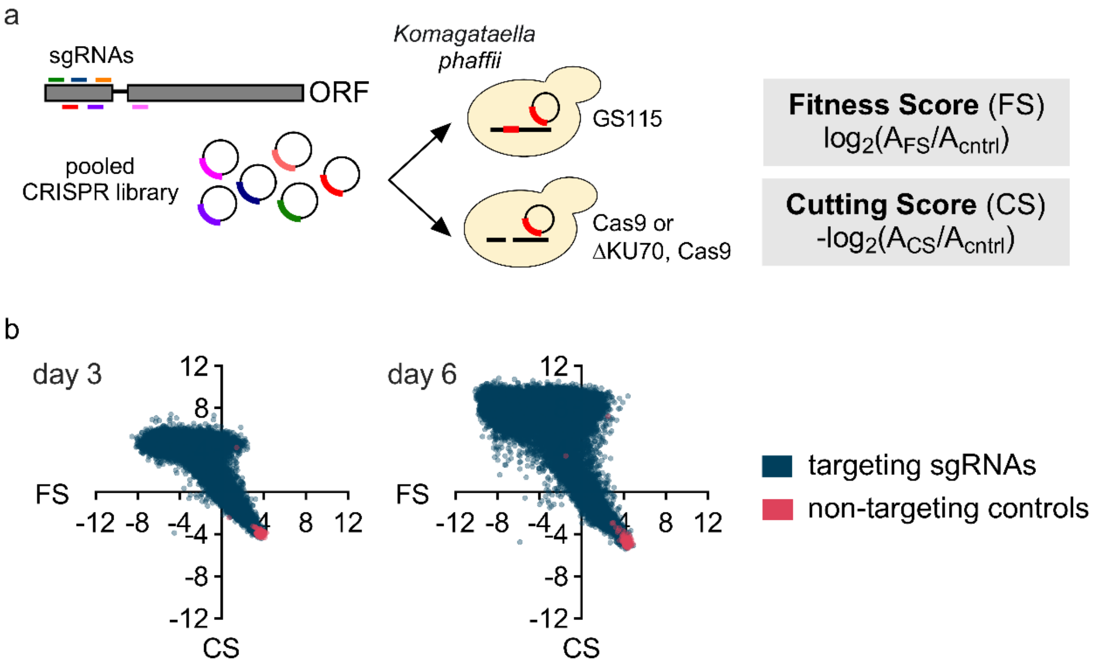


Fig. 3- 1. Genome-wide CRISPR-Cas9 single guide RNA (sgRNA) functional genetic screens in K. phaffii. **a)** Fitness and cutting score screens. *Komagataella phaffii* GS115 strain was used as the base strain for all experiments. GS115 *his4*::*CAS9* and GS115 *his4*::*CAS9 ΔKU70* strains were used for fitness score (FS) and cutting score (CS) experiments, respectively. GS115 and GS115 *ΔKU70* were used as the control strains for the FS and CS screens. A genome-wide sgRNA library was designed to target the first 300 bp of each expressed gene. The 6-fold coverage library was transformed into each strain and growth screens were performed to determine CS and FS for each sgRNA. **b)** Scatter plots of the CS and FS values generated on day 3 and 6 of the screens. Data points represent the average FS and CS values for triplicate experiments; each replicate was created with an independent library transformation.

3.3.2 In silico sgRNA design produces a highly active guide library

We designed a 6-fold genome-wide sgRNA library targeting 5309 protein coding sequences (CDSs) and 120 tRNAs in *K. phaffii* GS115 (**Fig. 3-2a** and **Supplementary Data 2**). The initial library included 169,034 sgRNAs targeting the first 300 bp of each CDS and tRNAs (**Supplementary Data 3**). Using a combined metric that accounted for the predicted activity of each guide and the uniqueness of each guide sequence, this large pool of guides was reduced to 31,634, including the top six ranked guides for each gene in the genome. An additional 350 non-targeting sgRNAs (randomly generated sequences with no homology to the GS115 genome, **Supplementary Data 4**) were added to the library for a total of 31,984 sgRNAs targeting 99.68% of the CDSs. Seventeen CDSs were excluded from the library due to the lack of unique guides (**Supplementary Table 3-1**).

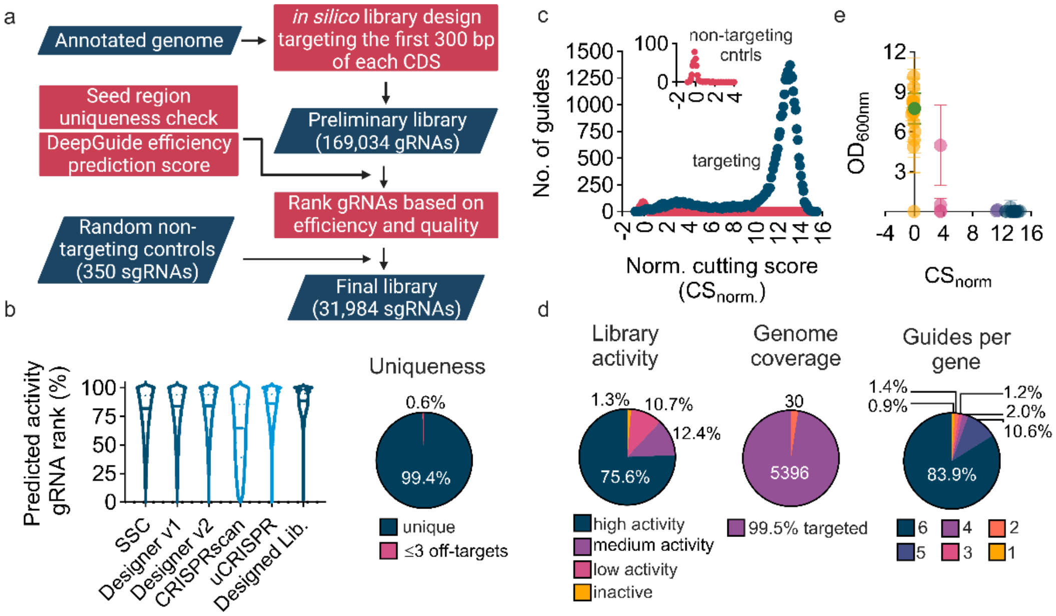


Fig. 3- 2. Genome-wide library design, genome-wide CS profile and validation. **a)** Schematic representation of the genome-wide sgRNA library design workflow. CHOPCHOP v3 and custom python scripts were used to identify all sgRNAs targeting the first 300 bp of each coding sequence (CDS) and tRNA genes. A series of guide activity prediction methods (five used by CHOPCHOP v3 plus DeepGuide[27] and a quality score (uniqueness and self-complementarity) were used to identify the best six sgRNAs targeting each gene. The final library consisted of 31,634 genome-targeting sgRNAs and 350 non-targeting controls. b) Criteria for choosing the best six sgRNAs for the final library. The violin plots show the ranked activity of guides as predicted by the five algorithms used by CHOPCHOP v3. 99.4% of the sgRNAs in the library are unique (see methods for uniqueness criteria) and only 0.6% of the library consists of sgRNAs with up to 3 off-targets and up to 3 mismatches (**Supplementary Table 3-7**). c) CS distribution on day 6. The CS for each sgRNA is normalized to the average CS of non-targeting controls. The presented CS values are the mean of three biological replicates. d) CS downstream analysis. 98.7% of the library consists of high (CSnorm > 11.46), medium (6.90 <CSnorm < 11.46), and low activity (1.36 <CSnorm < 6.90) sgRNAs. 392 sgRNAs were identified as inactive (CSnorm < 1.36). Active sgRNAs target 5396 genes in the GS115 genome, with only 30 genes not covered in the library. 83% of genes were targeted with 6 active sgRNAs. e) CS validation. 24 active (including highly active (dark blue), medium (purple), and low (magenta) activity) and 16 inactive (yellow) sgRNAs were chosen for validation experiments. sgRNAs were expressed in GS115 his4::CAS9 ΔKU70. Transformants with inactive sgRNAs showed growth similar to a control (green) strain, whereas cells transformed with active sgRNAs showed no or limited growth compared to control (3-day culture in SD-H, 2% glucose, 30 °C, 225 rpm). Data points and error bars represent the average of three biological replicates and one standard deviation, respectively.

Guides chosen to be in the final library are highly ranked by all activity predictors and over 99% have unique seed sequences (the 12 bp upstream of the PAM sequence) with no predicted off-target effects (**3-2b**, see "Materials and Methods" for more details). We focused our uniqueness criteria on the seed sequence because off-target effects have been shown to be more prominent with mismatches outside of the seed region and seed uniqueness is critical to on-target Cas9 effectiveness[29–33].

Using the designed library, we conducted a growth screen with cells containing disabled NHEJ to generate a CS profile across the genome (**Fig. 3-2c**, **Supplementary Fig. 3-1a**, and **Supplementary Data 1**). The CS of each sgRNA was normalized to the average CS of the non-targeting population (CS_{norm}). The CS distribution was found to be bimodal, with a large fraction of the library centered around a CS value of +13 compared to the non-targeting guide population (CS_{norm}). K-means clustering analysis classified the guides into four activity groups based on CS_{norm}: highly active ($CS_{norm} > 11.46$), medium activity ($6.90 < CS_{norm} < 11.46$), low activity ($1.36 < CS_{norm} < 6.90$), and inactive ($CS_{norm} < 1.36$) guides. Based on this analysis, only 1.3% of the guides in the library are inactive, while 75.6% are highly active (**Fig. 3-2d**). Active sgRNAs (including low, medium, and high activity) collectively targeted 5396 genes. Moreover, 83% of the genes were targeted by six active sgRNAs in the library, while 30 genes were not targeted by any active guide. Validation experiments on a subpopulation of guides confirmed that CS is an accurate representation of Cas9 activity (**Fig. 3-2e** and **Supplementary Fig. 3-2**). With one exception, Cas9-expressing NHEJ-deficient cells expressing twenty-four active sgRNAs exhibited either no or limited growth compared to the empty vector transformation ($p < 0.0005$). In contrast, 15 of 16 samples with inactive sgRNAs demonstrated growth comparable to the control, thus supporting CS as a quantitative metric for CRISPR-Cas9 activity. Taken together, the CS profiles, library analysis, and CS validation show that the designed library is highly active and has near complete genome-wide coverage of expressed genes.

3.3.3 Activity corrected fitness screens enable accurate essential gene classification

With the CS profile in-hand, we next set out to conduct a fitness screen and determine FS values for every guide in the library and gene in *K. phaffii* GS115 strain (Supplementary Data 1). The resulting library FS profile (FS values for every guide) was bimodal with distinctive peaks at FS approximately −2.8 and −8 (**Fig. 3-3a**). At earlier time points, the FS distribution was less pronounced (**Supplementary Fig. 3-1b**), therefore we used the day-6 time point to define essential genes under glucose growth conditions (2% glucose, SD-H, 30 °C). Our definition of gene essentiality, consistent with the established definition, includes both core essential genes indispensable for growth and conditionally essential genes that are related to the environmental context[34]. Using our acCRISPR analysis pipeline[15], low activity guides were analytically removed from the library before defining FS values per gene and calling essential genes. acCRSIPR identified a $CS_{threshold}$ of 7 to maximize library activity; only guides with a CS value of 7 or greater were used to calculate a gene's FS value (**Fig. 3-3b**). At this threshold, the library maintained an average CS value of 8.25, an average of 4.26 guides targeted each gene, and 1604 genes were classified as essential under given growth conditions (corrected p-value <0.05 per gene against a non-essential gene population, Supplementary Data 5). More than 99% of the predicted essential genes were targeted with more than one sgRNA, while genes with only one active sgRNA above the $CS_{threshold}$ were classified as low-confidence essential genes (**Supplementary Fig. 3-3**). In total, 1596 genes were classified as essential with high confidence.

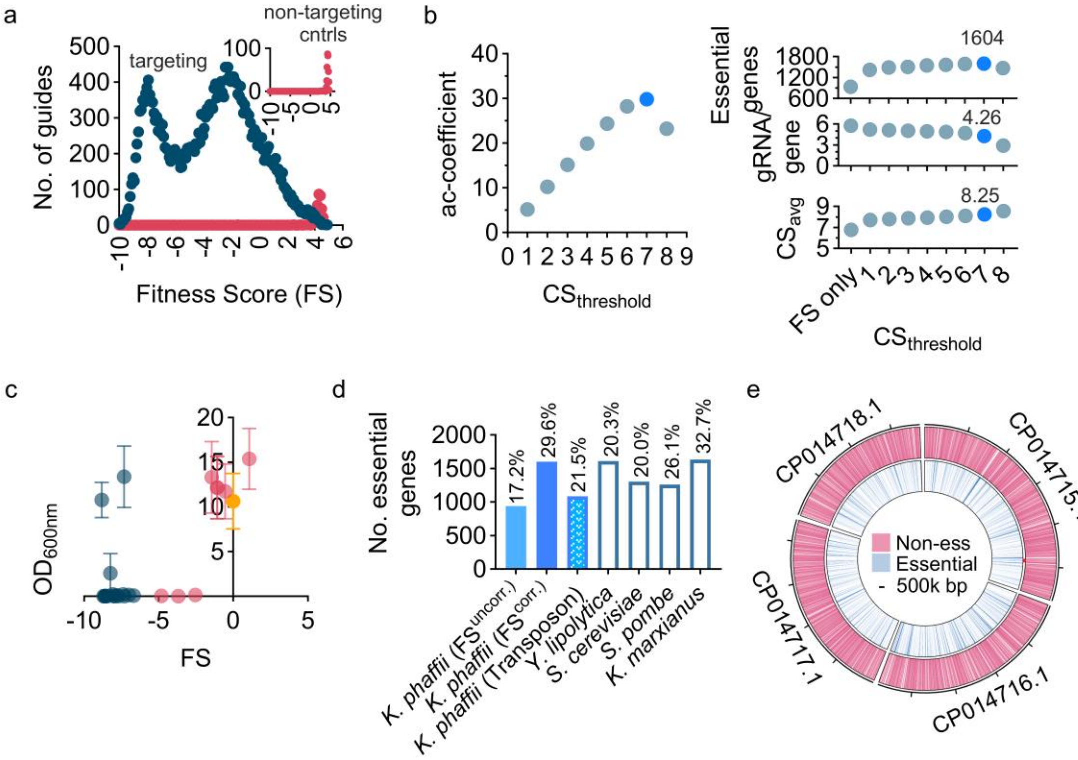


Fig. 3- 3. Activity corrected functional genetic screening in *K. phaffii*. **a)** FS frequency distribution per sgRNA. The presented FS values are the mean of three biological replicates per sgRNA at day 6. **b)** Essential gene identification using acCRISPR. The maximum activity correction coefficient (ac-coefficient) occurred at $CS_{threshold}$ value of 7, indicating the conditions for the highest library activity and coverage. At this threshold, 1604 genes were classified as essential (corrected p-value <0.05). Screens were conducted in SD-H, with 2% glucose, 30 °C. **c)** Individual validation of 17 predicted essential genes (dark blue) and 8 non-essential genes (magenta). A knockout in essential genes leads to low cell viability or cell death compared to a control (yellow). Data points and error bars represent the mean of three biological replicates and one standard deviation three days after subculturing in fresh selective media, respectively. **d)** The number of essential genes in *K. phaffii* with and without activity correction compared with essential gene calls from transposon analysis in *K. phaffii*[22], *Yarrowia lipolytica*[15], *Saccharomyces scerevisiae*[35], *Schizosaccharomyces pombe*[19], and *Kluyveromyces marxianus* (**Supplementary Data 6**). Values at the top of each bar represent the percentage of the total number of identified essential genes for each species/method. **e)** Distribution of predicted essential and non-essential genes in GS115's genome when grown on SD-H media with 2% glucose, 30 °C.

To validate the essentiality of the genes identified by acCRISPR, we selected 17 genes characterized as essential and 8 non-essential genes. Using one highly active guide per gene, we conducted a validation test similar to that conducted for CS validation; guides were transformed into GS115 *his4*::*CAS9* and allowed to grow for up to three days after transferring transformants to fresh selective media. Disruption of an essential gene should produce cultures with no-growth, while disruption of non-essential genes should have minimal effect on culture fitness. Of the 17 essential genes tested, 15 showed no or limited growth compared to the negative control ($p < 0.05$). Five of eight non-essential gene knockouts grew similar to the negative control (**Fig. 3-3c** and **Supplementary Fig. 3-4**), while three showed minimal or no growth.

Based on the analysis of model yeast species, roughly 20–30% of yeast genes are essential for growth. For example, 19.9% of S. cerevisiae genes are classified as essential[35], while in S. pombe an upward of 26.1% of genes are essential[19]. Our previous analysis of *Yarrowia lipolytica* identified 24.0% of genes as essential for growth on glucose[15], and a similar analysis of *Kluyveromyces marxianus* suggests that 30.8% of its genes are essential (**Fig. 3-3d** and **Supplementary Data 6**). Here, we make the comparison to these species as an additional validation step to the essential gene classification in *K. phaffi*i. Without activity correction via acCRISPR, only 934 *K. phaffii* genes (17.21% of all CDSs) were identified as essential, suggesting that including all guides in the library results in underestimation of gene essentiality. In addition, a genome-wide transposon insertion library, which is known to under-represent shorter genes[23], only identified 1086 essential genes in GS115 with high confidence and an additional 887 with low confidence[22]. The activity corrected screens conducted here classified a total of 1604 genes as essential (98.4% high confidence calls) or 29.55% of coding sequences in K. phaffii GS115, evenly distributed across the genome (**Fig. 3-3e).**

We further validated the essential gene set via Gene Ontology (GO) enrichment analysis (**Supplementary Fig. 3-5**)[36,37]. The analysis revealed multiple significantly enriched GO terms (adj. $p < 0.05$; see Supplementary Data 7 for all GO terms pertaining to molecular function (MF), biological process (BP), cellular component (CC), and KEGG pathway enrichment analysis) with markedly lower FS values compared to the average FS value of all genes. It was anticipated that terms functional for fundamental cell processes would be enriched. As expected, genes involved in translation, protein transport and maturation, DNA replication, ribosomal subunit export and assembly, and mitochondrial genes were significantly enriched. Taken together with the other validation methods described above, the essential genes identified from our CRISPR screens represent an accurate classification of essential genes.

3.3.4 Defining a consensus set of essential genes for *Komagataella phaffii* on glucose

The CRISPR-Cas9 screens conducted here, along with the transposon screen conducted by others, provide an opportunity to define a consensus set of essential genes for K. phaffii GS115 on glucose. Our validation experiments showed that the activity corrected CRISPR screen yielded a reasonably low false positive rate, but also identified the possibility of false negatives (**Supplementary Fig. 3-4**). Given this, we created a consensus set of essential genes by taking the union set called by both technologies (**Fig. 3-4a** and **Supplementary Data 8**). Among the 1086 high confidence essential genes characterized in the transposon study, 1064 have homologs based on the updated genome annotation used in our study[38]; only these genes were used to define the consensus set. The union set includes 1880 genes, 816 and 276 of which were only called by the CRISPR screen and the transposon study, respectively, and 788 genes called by both technologies.

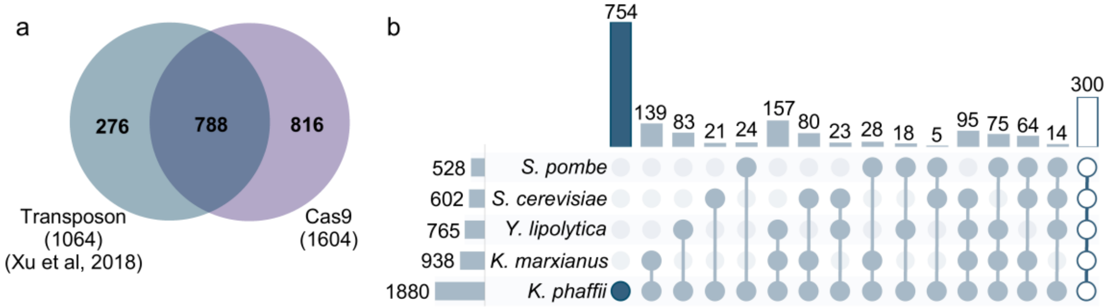


Fig. 3- 4. Identification of a consensus set of essential genes for *K. phaffii* **a)** Venn diagram representation of the number of essential genes identified based on our CRISPR-Cas9 screen, transposon analysis, and their overlap. The consensus essential gene list for K. phaffii GS115 on glucose is identified as the union of genes characterized as essential based on CRISPR-Cas9 screens and transposon analysis. **b)** Upset plot representation of the number of essential genes that are common between different yeast species. Values on the top of vertical bars represent the number of essential genes in K. phaffii that have essential homologs in other species. Values on the left of the horizontal bars are the intersection of essential genes between species.

The consensus set of 1880 essential genes for *K. phaffii* had 992, 765, 602, and 528 essential homologs in *K. marxianus*, *Y. lipolytica*, *S. cerevisiae* and *S. pombe*, respectively (**Fig. 3-4b**). Comparison between the consensus set and essential genes in other species also reveals a set of 268 core essential genes common to all five analyzed species as well as 760 genes exclusively essential to *K. phaffii*. Furthermore, non-essential genes in *K. phaffii* had homologs in various species: 1420 in *K. marxianus*, 1377 in *Y. lipolytica*, 1353 in *S. cerevisiae*, and 847 in *S. pombe*. BLAST analysis of these homologs identifies 350, 302, 202, and 184 genes that are essential in *K. marxianus*, *Y. lipolytica*, *S. cerevisiae*, and *S. pombe*, respectively. According to phylogenetic assessment (**Supplementary Fig. 3-6**), the divergence between *S. pombe* and *S. cerevisiae* occurred approximately 420 to 330 million years ago, leading to more genetic distinction among these two species[39]. *S. cerevisiae* and *K. phaffii* separated from each other more recently, around 250 million years ago[2]. This relatively more recent divergence likely accounts for the higher number of shared essential genes between *K. phaffii* and *S. cerevisiae* compared to *S. pombe*. *Y. lipolytica*, on the other hand, shares a common ancestor with *K. phaffii*[2], potentially contributing to the higher overlap in the number of common essential genes between *K. phaffii* and *Y. lipolytica* compared to the other analyzed species.

3.3.5 Gene Ontology enrichment analysis for essential genes

As additional analysis and validation, Gene Ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were conducted for the consensus set of essential genes, the core essential genes common between all five analyzed yeast species, and the essential genes solely belonging to *K. phaffii* (Supplementary Data 9 and 10). Enriched terms (adjusted *p*-value <0.05) for both analyses, GO terms and KEGG pathways, are represented in **Fig. 3-5, Fig. 3-6.**

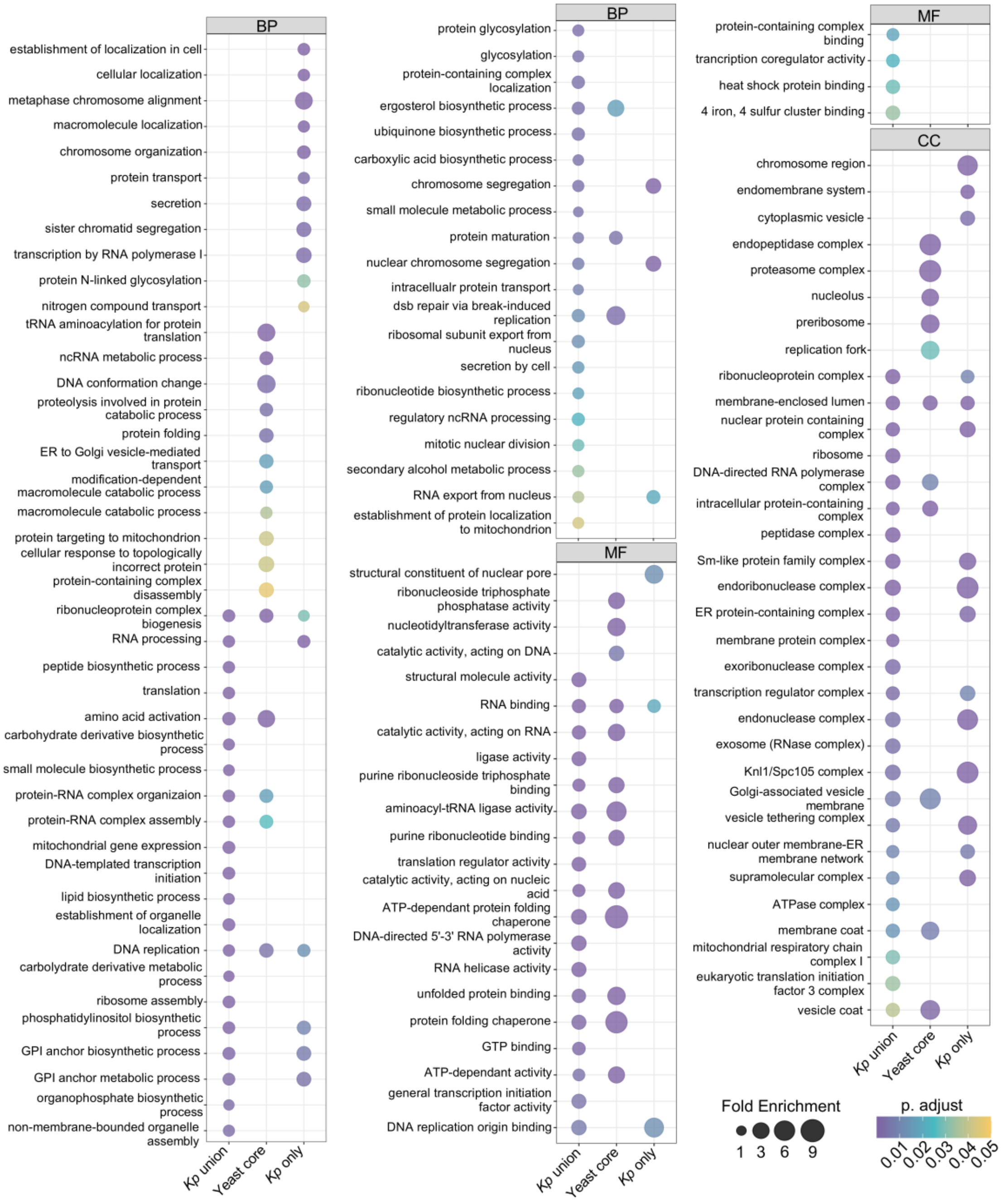


Fig. 3- 5. Comparison of functional profiles among different gene sets (Gene Ontology-enrichment analysis). Significantly enriched GO terms in biological process (BP), molecular function (MF), and cellular component (CC) categories (adjusted p-value <0.05) for the consensus set of essential genes (Kp union), the core essential genes between five analyzed yeast species (Yeast core), and essential genes solely belong to *K.*

phaffii (Kp only) are shown. Fold enrichment is the ratio of the frequency of input genes annotated in a term to the frequency of all genes annotated to that term.

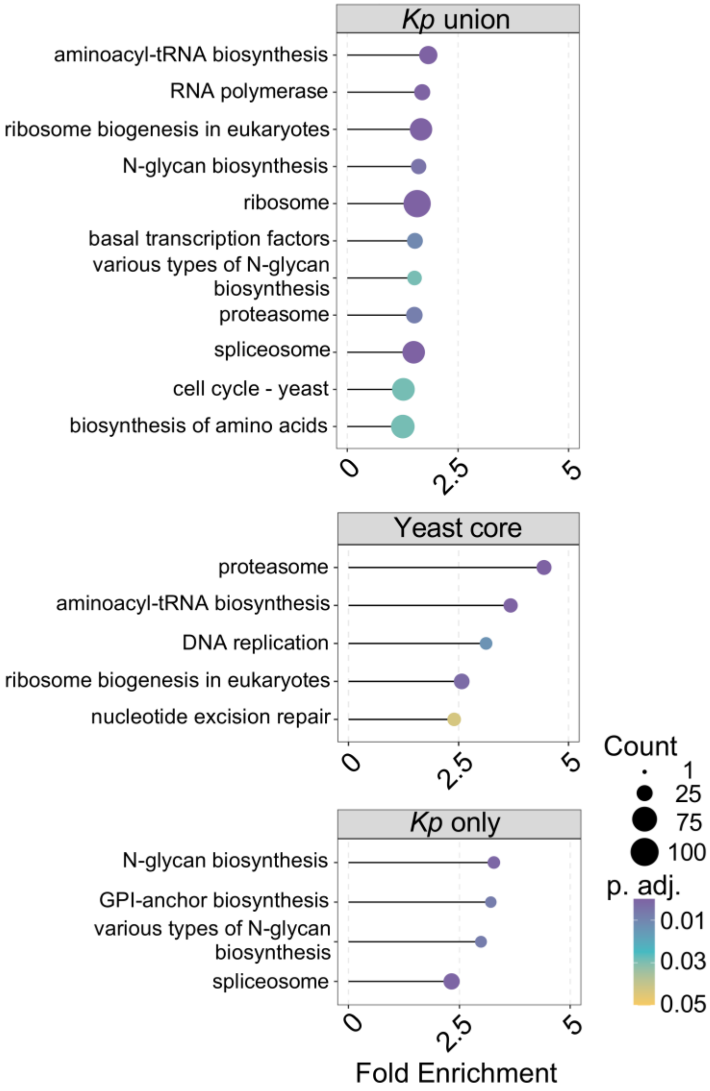


Fig. 3- 6. Comparison of functional profiles among different gene sets (Kyoto Encyclopedia of Genes and Genomes pathway analysis). Significantly enriched pathways (adjusted p-value <0.05) for the consensus set of essential genes (Kp union), core essential genes between five analyzed species (Yeast core), and essential genes solely belong to K. phaffii (Kp only). Count represents the number of genes annotated in a specific term, and fold enrichment is defined as the ratio of the frequency of input genes annotated in a term to the frequency of all genes annotated to that term.

As expected, in all three sets of essential genes (consensus, core, and *K. phaffii* specific) we identified vital cell processes and biological pathways (**Supplementary Figs. 3-7–9** and **Supplementary Data 11–13**). The general pattern is that there is minimal overlap of the enriched terms between the three datasets; only three of the 126 enriched GO terms are common between all three sets. These three GO terms are ribonucleoprotein complex biogenesis, DNA replication, and membrane-enclosed lumen. The yeast core enriched terms are most likely composed of conserved genes and pathways preserved through evolution between different species, e.g. tRNA aminoacylation for protein translation, DNA conformation change, protein folding, and cellular response to topologically incorrect proteins. However, terms specifically belonging to *K. phaffii* are composed of genes associated with unique, non-conventional characteristics of the "biotech yeast" and represent potential targets for metabolic engineering of this yeast.

One of the most important traits of *K. phaffii* is its capability to produce and secrete high titers of recombinant proteins. Various enriched GO terms exclusive to *K. phaffii* in the biological process (BP) category are related to protein production and secretion (**Fig. 3-5**). Protein transport, establishment of localization in cells, macromolecule localization, and mRNA export from nucleus are amongst the enriched GO terms only for *K. phaffii*. Multiple studies have shown overexpression of genes belonging to these terms to be associated with higher secretion of recombinant products. For instance co-overexpression of *S. cerevisiae* homologs of *SEC63* and *YDJ1* chaperones in *K. phaffii* were attributed to 7.6 times improvement of G-CSF secretion[40]. In addition, a study done with a *S. cerevisiae* strain with improved amylase production showed that *ERO1*, *BST1*, *SFB3*, *PEP5,* *SEC8,* and *EXO84* were upregulated, with all genes being involved in critical roles related to either protein folding or trafficking[41]. It was suggested that the observed upregulation in these genes might be an indication of the higher activity of the secretory pathway in this strain of *S. cerevisiae*. While these genes were not identified as essential in *S. cerevisiae*, they were categorized as essential in *K. phaffii* based on our screen.

Amongst the GO terms in cellular component (CC) category, endomembrane system, endoplasmic reticulum, and cytoplasmic vesicle are enriched in the *K. phaffii* only data set. Multiple studies have also demonstrated improved protein production with overexpression of genes belonging to these categories. For example, overexpressing the transcription factor *NRG1* in *K. phaffii* is associated with increases in the secretion of Fab2F5, recombinant human trypsinogen, and porcine trypsinogen[42]. Another *S. cerevisiae* study, showed that overexpression of the *LHS1* chaperone, which is involved in polypeptide translocation and folding in the ER lumen, increased shake-flask production levels of recombinant human serum albumin, granulocyte-macrophage colony-stimulating factor, and recombinant human transferrin[43]. Lastly, one study showed the influence of *ERO1* overexpression was able to increase nitrilase production in *K. phaffii*[44]. Given these examples, the identification of essential genes belonging to specific pathways via genome-wide knockout libraries can be used to add critical information to metabolic engineering design-build-test-learn cycles to engineer complex phenotypes such as secretion in which overexpression of essential genes can be beneficial.

Another industrially-relevant trait of *K. phaffii* is its ability to glycosylate recombinant proteins[45]. KEGG pathway enrichment analysis shows N-glycan and various types of N-glycan biosynthesis to be two of the significantly enriched pathways only in *K. phaffii* (**Fig. 3-6**). Fungi and mammals share initial steps in protein N-glycosylation, including site-specific transfer of a core oligosaccharide ($Glc_3Man_9GlcNAc_2$) to the nascent polypeptide. Downstream of the first glycosylation events, fungi exhibit a distinct processing pathway in comparison to mammalian cells. Fungi are limited to the addition of mannose and mannosylphosphate sugars to the glycoproteins, which leads to hyper-mannosylation (*S. cerevisiae*) or high-mannose structures (*K. phaffii*) of proteins causing immunogenicity in humans[46–48].

There are 31 genes associated with this pathway in the *K. phaffii* consensus set, among which 17 are exclusively essential in *K. phaffii* including both ER- (*SEC59*, *ALG5*, *ALG13*, *ALG3*, *ALG9*, *ALG12*, *ALG6*, *ALG8*, *OST1*, *OST3*, *SWP1*, *ROT2*, and *DFG10*) and Golgi-residing enzymes (*MNN2*, *MNN11*, MNN10, and *OCH1*). This gene set represents potential targets for metabolic engineering of non-native glycosylation patterns in *K. phaffii*. For instance, multiple studies have shown that endogenous *OCH1* knockout, a mannosyltransferase which initiates the first step of hypermannosylation in yeast, followed by introducing additional enzymes is crucial in humanizing the glycolysis pathway in *K. phaffii*[47,49,50]. While knocking out *OCH1* negatively impacts cell growth, as indicated by our CRISPR screen and other studies[51,52], the growth impediment is less pronounced in *K. phaffii* compared to *S.*

cerevisiae. We note that knockout of OCH1 leads to a serious impairment in growth and is called an essential gene in the CRISPR-Cas9 screens presented here.

Additional distinctive features of *K. phaffii* including the lack of one α-1,3-mannosyltransferase residing in the Golgi, leading to less hyperglycosylation, along with its mammalian-like stacked Golgi structure makes it a superior host for the production of glycoproteins compared to the conventional *S. cerevisiae* system[46–48].

Genome-wide knockout libraries thus enable the identification of crucial genes involved in biological pathways, facilitating the understanding of how these pathways differ between microorganisms and offer a novel tool in identification of gene targets to reverse engineer pathways in cells and for metabolic engineering.

3.4 Conclusion

High-throughput techniques play a crucial role in advancing metabolic engineering and driving forward genetics. However, these tools are not up to par for non-conventional hosts[53]. Here, we have addressed this issue by designing and validating a 6-fold coverage genome-wide sgRNA library composed of 30,848 active guides that target over 99% of protein coding sequences in the biotech yeast *Komagataella phaffii*. We also optimized the existing transformation protocols for this yeast, enabling the transformation of large-sized libraries for this host. Activity-validated sgRNA libraries can be used to improve screening accuracy, enhance genetic understanding, and aid in optimizing production strains. Notably, similar genome-wide sgRNA libraries have proven effective in finding hits to improve salt tolerance and uncover previously-unknown genes associated with lipid bio-production in *Yarrowia lipolytica*[15,25]. Application of this tool allowed us to define a consensus set of essential genes for this host on glucose. Through comparison with other yeasts, we have identified a set of essential genes exclusive to *K. phaffii*. These essential genes are promising candidates for overexpression, facilitating the engineering of complex phenotypes and advancing metabolic engineering efforts for *K. phaffii*.

3.5 Materials and Methods

3.5.1 Strains and culture conditions

Komagataella phaffii GS115 (Invitrogen), a strain with histidine deficiency, was used for all experiments (**Supplementary Table 3-2**). GS115 *his4*::*CAS9* strain was constructed by integrating p*ENO1*-Cas9-PptefT expression cassette into cells' knocked-out *HIS4* loci. GS115 *ΔKU70* and GS115 *his4*::*CAS9* *ΔKU70* strains were constructed by disrupting *KU70* using CRISPR-Cas9.

All yeast culturing was done at 30 °C in 14 ml polypropylene tubes or in 2 L baffled flasks as noted, at 225 rpm. Under non-selective conditions, yeast strains were initially grown in YPD (1% Bacto yeast extract, 2% Bacto peptone, 2% glucose). Cells were transformed with plasmids expressing sgRNAs and transformants were recovered in histidine-deficient media (SD-his; 0.67% Difco yeast nitrogen base without amino acids, 0.069% CSM-his (Sunrise Science, San Diego, CA), and 2% glucose).

3.5.2 Plasmid construction

All plasmid construction and propagation were conducted in *E. coli* TOP10. Cultures were conducted in Luria-Bertani (LB) broth with either 100 mg/L ampicillin or 50 mg/L kanamycin 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.

All plasmids, primers, and sgRNAs used in this work are listed in **Supplementary Table 3-3 to Supplementary Table 3-5.** The D-227 vector containing *CAS9* and a gRNA expression cassette was a kind donation from the Love lab at Massachusetts Institute of Technology[10]. For integration of *CAS9* into *K. phaffii*'s genome, first a highly active *HIS4*-targeting sgRNA was cloned into D-227 vector by digesting the plasmid with BbVCI enzyme (NEB) according to the manufacturer's instructions. gRNA cloning was carried out according to a protocol developed in the lab previously[54]. Primers for sgRNA cloning were obtained from Integrated DNA Technology (IDT). Successful cloning of the sgRNA fragment was confirmed by Sanger sequencing. Next, 1000 bp directly upstream and downstream of the *HIS4*-targeting sgRNA on the genome was PCR amplified and cloned on the upstream and downstream of the p*ENO1*-Cas9-PptefT on D-227 vector using New England BioLabs (NEB) NEBuilder® HiFi DNA Assembly Master Mix. This plasmid was transformed into GS115. *CAS9* integration was verified with PCR amplification of the *HIS4* loci and Sanger sequencing. For all the PCR amplifications in this study Q5 high fidelity polymerase (NEB) was used according to the manufacturer's instructions.

The backbone of the sgRNA library plasmid (pCRISPRpp) was constructed by PCR amplification of the PARS1 sequence from *K. phaffii*'s genome[9]. The *E. coli* origin of replication and ampicillin resistance gene were PCR amplified from pCRISPRyl (Addgene #70007)[55]. CYC1t and pTEF1 were both PCR amplified from

BB3cK_pGAP_23*_pTEF_Cas9 (Addgene #104909)[56]. sgRNA expression cassette (ptRNA1_tRNA1_tracrRNA) was PCR amplified from D-227 plasmid. Pp*HIS4* gene was PCR amplified from pMJA089 (Addgene #128518)[57]. All fragments were cloned to each other to make a single plasmid with NEBuilder® HiFi DNA Assembly Master Mix.

3.5.3 gRNA library design

CHOPCHOP v3 [33]was used to design the sgRNA library for *K. phaffii*. The GS115 reference genome and annotation was downloaded from RefSeq at NCBI (sequence assembly version ASM174695v1, RefSeq assembly accession: GCA_001746955.1) and Bioproject PRJNA66950[38]. sgRNAs were designed to target the first 300 bp of each coding sequence and tRNA genes to maximize a functional knockout in the gene in case of a CRISPR-induced indel. CHOPCHOP v3 was used to design a preliminary library of 169,034 sgRNAs (Supplementary Data 3). Each guide within this preliminary library was characterized by multiple targeting efficiency predictive parameters from various tools including Designer v1[58], Designer v2[59], CRISPRscan[60], SSC[61], and uCRISPR[62]. To enhance the library design process, we also introduced a CS prediction score identified from DeepGuide[27] trained based on *Yarrowia lipolytica* PO1f CRISPR-Cas9 genome-wide sgRNA library CS data. A naive score for each sgRNA was calculated as the aggregate of all the aforementioned normalized targeting efficiency scores.

The uniqueness of each 20 bp sgRNA was analyzed with CHOPCHOP v3 built-in MM0, MM1, MM2, and MM3 scores determining the number of off-target transcripts for each sgRNA with 0, 1, 2, and 3 mismatches, respectively. We also incorporated an extra measure of uniqueness, Seed_MM0, identifying the number of sgRNAs targeting anywhere within the genome with 0 mismatches in the seed region, the last 12 bp of the sgRNA immediately preceding the NGG PAM motif-compared to our sgRNA of interest. Numerous studies have documented that the uniqueness of this seed sequence is a pivotal factor in minimizing the off-target effects of Cas9[29–31]. Additionally, a self-complementarity score was employed to predict the likelihood of the sgRNA forming a secondary structure with itself, potentially reducing the targeting efficiency[33]. A comprehensive quality score was assigned to each sgRNA taking into account all uniqueness and self-complementarity scores. A quality score of 1 signifies an sgRNA that not only possesses uniqueness in both its 20 bp sequence and seed region but also exhibits a minimal likelihood of forming secondary structures. The detailed breakdown of all the defined quality scores can be found in **Supplementary Table 3-6**.

sgRNAs designed for each coding sequence were initially ranked based on their "quality" score and then the top sgRNAs with the highest "naive" score were chosen for the final library for each coding sequence or tRNA gene. Over 99% of the sgRNAs in the library had a quality score of 1 (**Supplementary Table 3-7**). Three hundred and fifty sgRNAs with random sequences were also included as non-targeting controls (Supplementary Data 4). All designed sgRNAs along with additional data are available in Supplementary Data 2.

3.5.4 sgRNA library cloning

60mer linkers were added 5′ and 3′ of each designed sgRNA enabling assembly into pCRISPRpp (**Supplementary Table 3-4**) and obtained as a pooled oligonucleotide library (Twist BioScience, CA, USA). The library was amplified for 9 cycles with Kapa polymerase (Roche) using the oligonucleotides 5′ tagtggtagaaccaccgcttgtc and 5′ actttttcaagttgataacggactagcc and assembled into pCRISPRpp linearized by BbvCI digestion, and dephosphorylated with quick CIP, using the NEBuilder Hi-Fi Assembly kit (New England Biolabs). To ensure representation of all variants in the population, >330,000 colonies were obtained (*i.e.*, >10 colonies per sgRNA), and the library validated by insert PCR amplification with the oligonucleotides 5′ agccaatcctactacattgatccg and 5’ gtcatgataataatggtttcttagacg. The amplicon library was sequenced on the Illumina MiSeq platform and the data analyzed using custom library quality control pipelines (**Supplementary Data 14**).

3.5.5 Yeast transformation and screening

Transformation of *K. phaffii* was done using a previously described method, with slight modifications[63]. Two mL of YPD was inoculated with a single colony of the strain of interest and grown in a 14 mL tube with shaking at 225 rpm overnight. 4 × 10^7 cells were transferred to 150 ml of YPD in a 500 ml baffled shake flask and grown for ∼14 h (until the culture reached a final OD_{600} = 1.8). 100 ml of cells were chilled on ice for 1.5 h, washed with 1 M ice-cold sorbitol three times, incubated with 25 ml of pretreatment solution (0.1 M lithium acetate (LiAc), 30 mM dithiothreitol (DTT), 0.6 M sorbitol, and 10 mM tris-HCl pH = 7.5) for 30 min at room temperature, and washed three more times with 1 M ice-cold sorbitol. For each transformation, 8 × 10^8 cells were mixed with 1 µg of library to a final volume of 80 µl, incubated on ice for 15 min, and pulsed at 1.5 kV with Bio-Rad MicroPulser Electroporator in an ice-cold 0.2-cm-gap cuvette. Immediately after electroporation shock, 1 ml ice-cold solution of YPD and 1 M sorbitol was added to each cuvette. Cells were transferred to 1 ml YPD and 1 M sorbitol in 14 ml tubes, incubated for 3 h at 30 °C and 225 rpm for recovery, washed with 1 ml of room-temperature autoclaved water to get rid of the excess plasmid DNA in samples, and transferred to selective media. All centrifugation was done at 3000×*g* for 5 min at 4 °C. The relationship between cell number and OD_{600} was calculated according to 1 OD_{600} = 5 × 10^7 cells/mL.

For library transformations, 10 separate transformations were pooled together after recovery to maintain library representation (100-fold coverage, total transformants per biological replicate). Pooled transformants were transferred to 750 ml SD-his for outgrowth experiments in a 2 L baffled shake flask. Three biological replicates were performed for each strain. Transformation efficiency for each replicate and strain is presented in **Supplementary Table 3-8**. Cells reached to confluency after 3 days (OD_{600} ≈ 8). 1 ml of cells were transferred to 50 ml fresh SD-his in 250 ml baffled shake flasks to perform outgrowth experiments and were allowed to grow for three more days. The experiment was stopped after reaching confluency again on day six of the screen. At each time point, 1 ml of culture was stored at −80 °C to isolate sgRNA expression plasmids for deep sequencing.

3.5.6 Library isolation and sequencing

Frozen culture samples from pooled screens were thawed. Plasmids were isolated from each sample using a Zymo Yeast Plasmid Miniprep Kit (Zymo Research). 500 µL of each sample was divided into two tubes to account for the capacity of the yeast miniprep kit, specifically to ensure complete lysis of the cells using Zymolyase. The split miniprepped samples from a single strain and replicate were pooled again, and the plasmid copy number was quantified using quantitative PCR with qPCR_GW.F, qPCR_GW.R, and SsoAdvanced Universal SYBR Green Supermix (Bio-Rad). Each pooled sample was confirmed to contain at least 10^7 plasmids ensuring sufficient coverage of the sgRNA library. Recovered plasmid copy number and coverage for each sample and replicate is presented in **Supplementary Table 3-9**.

To prepare samples for next generation sequencing (NGS), isolated plasmids from each sample were used as PCR templates using forward (NGS1-4.F) and reverse primers (NGS1-9.R). Different forward and reverse barcodes and pseudo-barcodes were used in primers to increase complexity for NGS and to enable us to differentiate between samples later on. NGS primers were ordered as Ultramer DNA oligos from IDT. At least 0.5 ng of the recovered plasmids (~molecules) were used to amplify the amplicons in a 16-cycle PCR reaction to minimize any bias. PCR products were cleaned by a double-sided cleanup technique using AMPure XP beads and tested with a Bioanalyzer to ensure the correct length of amplicons. 80 nmol of FS and cutting score CS samples were pooled together separately and submitted for sequencing on a NextSeq2000 using a P2 100 cycle kit.

3.5.7 Cutting and fitness score calculations

Based on our acCRISPR analysis pipeline[15], fitness score (FS) and cutting score (CS) was calculated by first adding a pseudo-count of one to each raw count before normalization. The read counts for each sgRNA were normalized to the total number of reads for that specific sample. Fitness score value for each sgRNA was calculated as the $\log_2$ ratio of normalized read counts obtained in GS115 *his4*::*CAS9* to normalized counts in GS115 strain. Similarly, cutting score (CS) was defined as the -$\log_2$ ratio of normalized reads obtained in GS115 *his4*::*CAS9 ΔKU70* to counts in GS115 *ΔKU70* (**Supplementary Data 1**).

3.5.8 Essential gene identification

FS and CS values of sgRNA at day six were used as input to acCRISPR v1.0.0[15] to identify essential genes from the screen. A $CS_{threshold}$ of 7.0 was used to remove low-activity sgRNAs from the original library, due to the maximum value of ac-coefficient at this threshold. FS of a gene was computed by acCRISPR as the average of FS of all sgRNAs with CS above 7.0 targeting that gene. Genes having FDR-corrected $p < 0.05$ were deemed as essential.

3.5.9 Finding essential gene homologs in S. cerevisiae, S. pombe, Y. lipolytica, and K. marxianus

Sequences of genes essential identified in this study and/or in the transposon study[22] were aligned to genes in *S. cerevisiae*, *S. pombe,* and *Y. lipolytica* using BLASTP, and to genes in *K. marxianus* using TBLASTN. *S. cerevisiae* essential genes (phenotype:inviable) were retrieved from the *Saccharomyces* Genome Database (SGD), *S. pombe* essential genes were taken from[19], and *Y. lipolytica* essential genes were taken from the consensus set defined in[15]. *K. marxianus* essential genes were identified from a CRISPR-Cas9 genome-wide library in the CBS6556 strain in our lab. Pairs of query and subject sequences having >40% identity from BLAST were deemed as homologs.

3.5.10 Experimental validation of fitness and cutting scores

Selected genes/sgRNAs were chosen for essential gene/cutting score validations, respectively. The essential gene validations were done by performing a single-gene knockout using high cutting score sgRNAs targeting 17 predicted essential genes and 8 non-essential genes in the GS115 *his4*::*CAS9* background. Additionally, sixteen inactive ($CS_{norm} < 1.36$), four low-activity ($1.36 < CS_{norm} < 6.90$), two medium-activity ($6.90 < CS_{norm} < 11.46$), and sixteen high-activity ($CS_{norm} > 11.46$) sgRNAs were chosen for CS validations in the GS115 *his4*::*CAS9 ku70* background. Individual plasmids containing sgRNAs were cloned as was mentioned previously. Transformants were grown in 4 mL of SD-his for two days, followed by sub-culturing in 2 mL of fresh selective media. The OD_{600} of the samples were measured three days after the sub-culture. All of the validation experiments were done in three biological replicates.

3.5.11 Functional enrichment analysis

The organism package for *K. phaffii* GS115 (NCBI Taxonomy ID: 644223) was created using the AnnotationForge package (version 1.44.0)[64]. ClusterProfiler package (version 4.10.0) was used for functional enrichment analysis[65,66]. GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analysis were applied to the genes in the consensus set, yeast core set, and *K. phaffii*-specific set. The *p*-value was calculated by Fisher's exact test. Benjamini-Hochberg procedure was applied to correct *p*-values. Significant GO terms and pathways were identified with a cutoff for adjusted *p*-value (adj. *p*-value <0.05). Fold enrichments, defined as the ratio of the frequency of input genes annotated in a term to the frequency of all genes annotated to that term, for all the enriched terms were also calculated to interpret the results better. To get a more effective interpretation from the analysis, some redundant GO terms (with semantic similarities over 0.7) were removed by applying the *simplify* function in the ClusterProfiler package (version 4.10.0). For some GO terms with a parent-child semantic relationship having the same *p*-values and geneRatio (ratio of input genes annotated in a term), the parent terms were eliminated from the list.

3.6 Code availability

Source code for the CRISPR-Cas9 library design can be found at https://github.com/ianwheeldon/Kphaffii_library_design.git/. Custom python scripts that were used for the processing of Illumina reads to generate sgRNA abundance for the Cas9 screens can also be found at the same link.

3.7 CRediT authorship contribution statement

Aida Tafrishi: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Varun Trivedi:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zenan Xing:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis. **Mengwan Li:** Investigation, Data curation. **Ritesh Mewalal:** Writing – review & editing, Methodology, Investigation. **Sean R. Cutler:** Writing – review & editing, Supervision, Methodology. **Ian Blaby:** Supervision, Methodology, Investigation. **Ian Wheeldon:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

3.8 Acknowledgments

This work was supported by NSF-1951942, NSF-2225878, and NSF-1922642. The work (proposal: https://doi.org/10.46936/10.25585/60001310) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. We thank Dipankar Baisya for his assistance with DeepGuide and Adithya Ramesh for guidance with the library design. We also thank Dr. J. Christopher Love for providing the Cas9 expressing plasmid, D-227.

References

1. Ahmad, M., Hirz, M., Pichler, H. & Schwab, H. Protein expression in Pichia pastoris: recent achievements and perspectives for heterologous protein production. *Appl. Microbiol. Biotechnol.* **98**, 5301–5317 (2014).
2. Bernauer, L., Radkohl, A., Lehmayer, L. G. K. & Emmerstorfer-Augustin, A. Komagataella phaffii as Emerging Model Organism in Fundamental Research. *Front. Microbiol.* **11**, 607028 (2020).
3. Daly, R. & Hearn, M. T. W. Expression of heterologous proteins in Pichia pastoris: a useful experimental tool in protein engineering and production. *J. Mol. Recognit.* **18**, 119–138 (2005).
4. Gasser, B. *et al.* Pichia pastoris: protein production host and model organism for biomedical research. *Future Microbiol.* **8**, 191–208 (2013).
5. Ata, Ö. *et al.* What makes Komagataella phaffii non-conventional? *FEMS Yeast Res.* **21**, (2021).
6. Love, K. R., Dalvie, N. C. & Love, J. C. The yeast stands alone: the future of protein biologic production. *Curr. Opin. Biotechnol.* **53**, 50–58 (2018).
7. Cai, P. *et al.* Recombination machinery engineering facilitates metabolic engineering of the industrial yeast Pichia pastoris. *Nucleic Acids Res.* **49**, 7791–7805 (2021).
8. Weninger, A., Hatzl, A.-M., Schmid, C., Vogl, T. & Glieder, A. Combinatorial optimization of CRISPR/Cas9 expression enables precision genome engineering in the methylotrophic yeast Pichia pastoris. *J. Biotechnol.* **235**, 139–149 (2016).
9. Cregg, J. M., Barringer, K. J., Hessler, A. Y. & Madden, K. R. Pichia pastoris as a host system for transformations. *Mol. Cell. Biol.* **5**, 3376–3385 (1985).
10. Dalvie, N. C. *et al.* Host-Informed Expression of CRISPR Guide RNA for Genomic Engineering in Komagataella phaffii. *ACS Synth. Biol.* **9**, 26–35 (2020).
11. Liu, R., Chen, L., Jiang, Y., Zhou, Z. & Zou, G. Efficient genome editing in filamentous fungus Trichoderma reesei using the CRISPR/Cas9 system. *Cell Discov* **1**, 15007 (2015).
12. Löbs, A.-K., Engel, R., Schwartz, C., Flores, A. & Wheeldon, I. CRISPR-Cas9-enabled genetic disruptions for understanding ethanol and ethyl acetate biosynthesis in Kluyveromyces marxianus. *Biotechnol. Biofuels* **10**, 164 (2017).
13. Dong, C. *et al.* A genome-wide CRISPR-Cas9 knockout screen identifies essential and growth-restricting genes in human trophoblast stem cells. *Nat. Commun.* **13**, 2548 (2022).

14. 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).

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

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

17. Trivedi, V., Ramesh, A. & Wheeldon, I. Analyzing CRISPR screens in non-conventional microbes. *J. Ind. Microbiol. Biotechnol.* **50**, (2023).

18. Guo, Y. *et al.* Integration profiling of gene function with dense maps of transposon integration. *Genetics* **195**, 599–609 (2013).

19. 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).

20. Michel, A. H. *et al.* Functional mapping of yeast genomes by saturated transposition. *Elife* **6**, (2017).

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

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

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

24. Morgens, D. W., Deans, R. M., Li, A. & Bassik, M. C. Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. *Nat. Biotechnol.* **34**, 634–636 (2016).

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

26. Zhang, G., Luo, Y., Dai, X. & Dai, Z. Benchmarking deep learning methods for predicting CRISPR/Cas9 sgRNA on- and off-target activities. *Brief. Bioinform.* **24**, (2023).

27. 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).

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

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

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

31. 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).

32. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821 (2012).

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

34. Bosch-Guiteras, N. & van Leeuwen, J. Exploring conditional gene essentiality through systems genetics approaches in yeast. *Curr. Opin. Genet. Dev.* **76**, 101963 (2022).

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

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

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

38. Alva, T. R., Riera, M. & Chartron, J. W. Translational landscape and protein biogenesis demands of the early secretory pathway in Komagataella phaffii. *Microb. Cell Fact.* **20**, 19 (2021).

39. Sipiczki, M. Where does fission yeast sit on the tree of life? *Genome Biol.* **1**, REVIEWS1011 (2000).

40. Zhang, W. *et al.* Enhanced secretion of heterologous proteins in Pichia pastoris following overexpression of Saccharomyces cerevisiae chaperone proteins. *Biotechnol. Prog.* **22**, 1090–1095 (2006).

41. Liu, Z. *et al.* Improved production of a heterologous amylase in Saccharomyces cerevisiae by inverse metabolic engineering. *Appl. Environ. Microbiol.* **80**, 5542–5550 (2014).

42. Stadlmayr, G., Benakovitsch, K., Gasser, B., Mattanovich, D. & Sauer, M. Genome-scale analysis of library sorting (GALibSo): Isolation of secretion enhancing factors for recombinant protein production in Pichia pastoris. *Biotechnol. Bioeng.* **105**, 543–555 (2010).

43. Payne, T. *et al.* Modulation of chaperone gene expression in mutagenized Saccharomyces cerevisiae strains developed for recombinant human albumin

production results in increased production of multiple heterologous proteins. *Appl. Environ. Microbiol.* **74**, 7759–7766 (2008).

44. Shen, Q. *et al.* Engineering a Pichia pastoris nitrilase whole cell catalyst through the increased nitrilase gene copy number and co-expressing of ER oxidoreductin 1. *Appl. Microbiol. Biotechnol.* **104**, 2489–2500 (2020).
45. Barone, G. D. *et al.* Industrial Production of Proteins with Pichia pastoris-Komagataella phaffii. *Biomolecules* **13**, (2023).
46. De Pourcq, K., De Schutter, K. & Callewaert, N. Engineering of glycosylation in yeast and other fungi: current state and perspectives. *Appl. Microbiol. Biotechnol.* **87**, 1617–1631 (2010).
47. Hamilton, S. R. *et al.* Production of complex human glycoproteins in yeast. *Science* **301**, 1244–1246 (2003).
48. Wildt, S. & Gerngross, T. U. The humanization of N-glycosylation pathways in yeast. *Nat. Rev. Microbiol.* **3**, 119–128 (2005).
49. Choi, B.-K. *et al.* Use of combinatorial genetic libraries to humanize N-linked glycosylation in the yeast *Pichia pastoris*. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 5022–5027 (2003).
50. Jacobs, P. P., Geysens, S., Vervecken, W., Contreras, R. & Callewaert, N. Engineering complex-type N-glycosylation in Pichia pastoris using GlycoSwitch technology. *Nat. Protoc.* **4**, 58–70 (2009).
51. Dalvie, N. C., Lorgeree, T., Biedermann, A. M., Love, K. R. & Love, J. C. Simplified Gene Knockout by CRISPR-Cas9-Induced Homologous Recombination. *ACS Synth. Biol.* **11**, 497–501 (2022).
52. Moser, J. W., Wilson, I. B. H. & Dragosits, M. The adaptive landscape of wildtype and glycosylation-deficient populations of the industrial yeast Pichia pastoris. *BMC Genomics* **18**, 597 (2017).
53. 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).
54. Schwartz, C. & Wheeldon, I. CRISPR-Cas9-Mediated Genome Editing and Transcriptional Control in Yarrowia lipolytica. *Methods Mol. Biol.* **1772**, 327–345 (2018).
55. 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).
56. Gassler, T., Heistinger, L., Mattanovich, D., Gasser, B. & Prielhofer, R. CRISPR/Cas9-Mediated Homology-Directed Genome Editing in Pichia pastoris. *Methods Mol. Biol.* **1923**, 211–225 (2019).

57. Yang, J. *et al.* Hygromycin-resistance vectors for gene expression in Pichia pastoris. *Yeast* **31**, 115–125 (2014).

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

59. 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).

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

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

62. 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).

63. Wu, S. & Letchworth, G. J. High efficiency transformation by electroporation of Pichia pastoris pretreated with lithium acetate and dithiothreitol. *Biotechniques* **36**, 152–154 (2004).

64. Marc Carlson, H. P. *AnnotationForge*. (Bioconductor, 2017). doi:10.18129/B9.BIOC.ANNOTATIONFORGE.

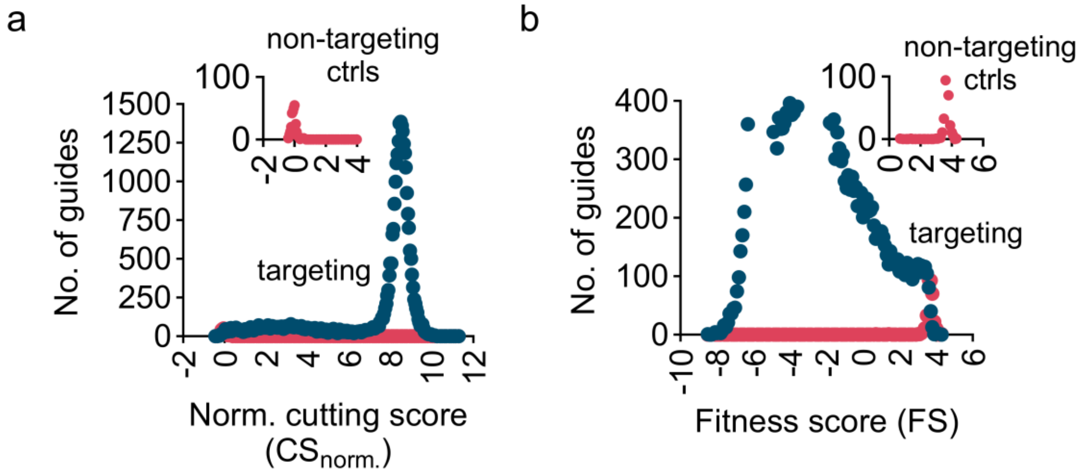
65. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).

66. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

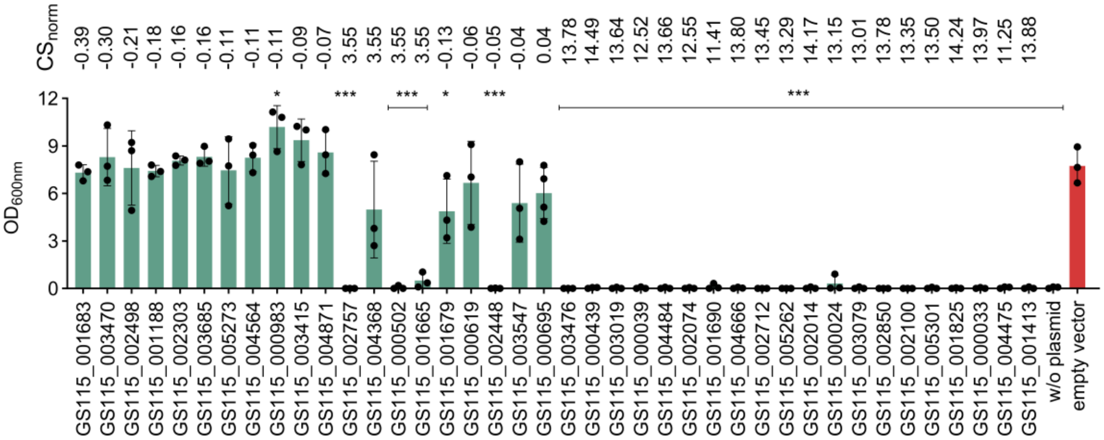
67. Carlson, M. & Pages, H. AnnotationForge: tools for building SQLite-based annotation data packages. R Packag. version 1.32. 0. Preprint at (2020).

3.10 Supplementary information

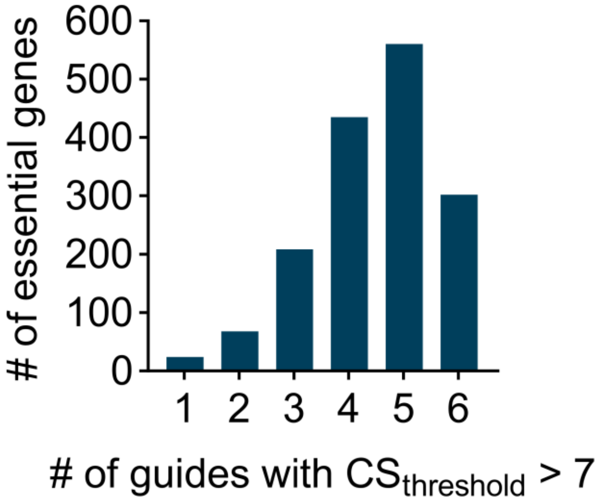
3.10.1 Supplementary figures



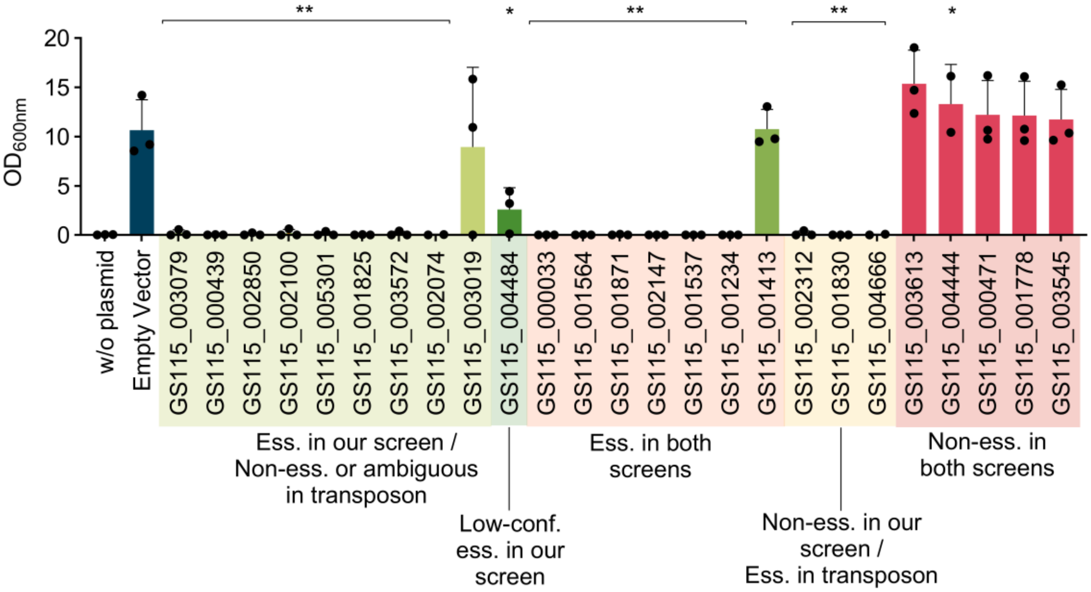
Supplementary Fig. 3- 1. Frequency distribution of **a)** normalized cutting score (CS); and **b)** fitness score (FS) data on day 3. The presented CS and FS values are the mean of three biological replicates. The presented CS for each sgRNA is normalized to the average CS of non-targeting controls.



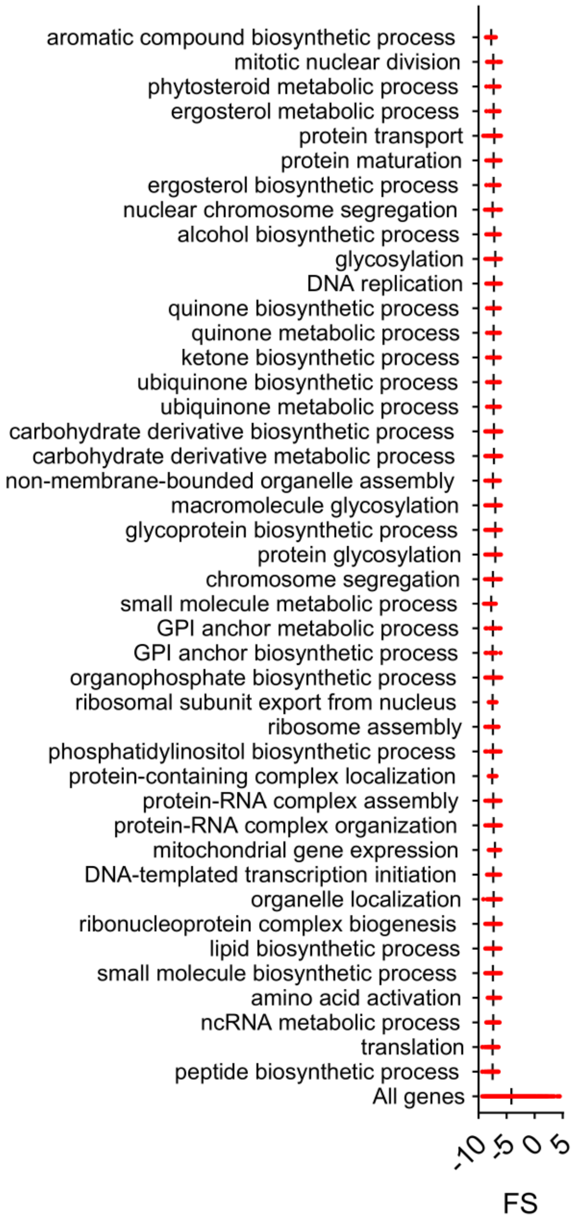
Supplementary Fig. 3- 2. Experimental CS validations obtained from CRISPR-Cas9 screen. Final OD600 of Cas9-expressing NHEJ-deficient cells expressing 24 high CS and 16 low CS sgRNAs. Transformants were grown in SD-H for two days right after electroporation, followed by subculturing in fresh SD-H media and were allowed to grow for three more days. Cells were also transformed with an empty vector as the control to show the impact of the presence of sgRNA. GS115 containing no plasmid was used as the negative control, showing no growth post electroporation. Validation experiments were done in three biological replicates. The bars show the mean of the replicates, data points represent OD of each individual replicate, and the error bars represent one standard deviation. * $p < 0.05$, *** $p < 0.0005$; one-tailed unpaired t-test)



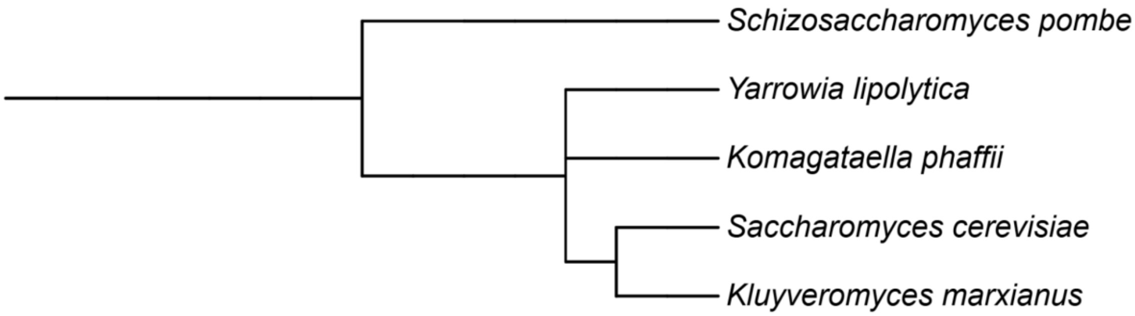
Supplementary Fig. 3- 3. Predicted essential gene coverage with activity-validated sgRNAs with a $CS_{threshold} > 7$. More than 98% of the predicted essential genes are covered with more than one highly active sgRNA further validating the essential gene identification.



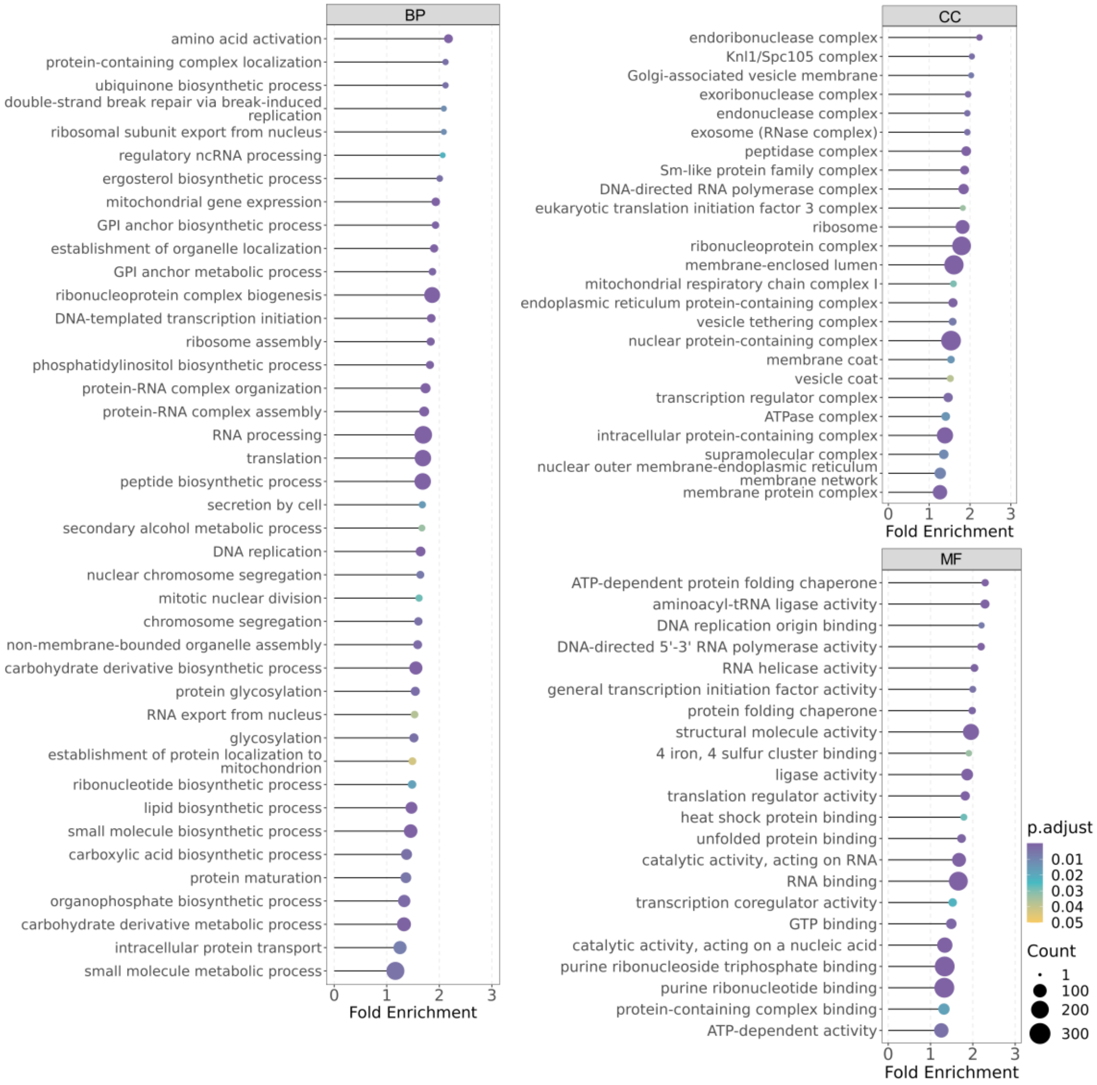
Supplementary Fig. 3- 4. Experimental validation of CRISPR-Cas9 essential and non-essential genes from acCRISPR analysis. Final OD of Cas9 expressing cells expressing sgRNAs targeting essential, low-confidence essential, and non-essential genes. Transformants were grown in SD-H for two days right after electroporation, followed by subculturing in fresh SD-H media and were allowed to grow for three more days. Cells were also transformed with an empty vector as the control to show the impact of gene essentiality on cell fitness. GS115 containing no plasmid was used as the negative control, showing no growth post electroporation. Validation experiments were done in three biological replicates. The bars show the mean of the replicates, data points represent OD of each individual replicate, and the error bars represent one standard deviation. Statistical analysis is done between transformants with sgRNA-containing plasmids and the empty vector. * $p < 0.05$, ** $p < 0.005$; one-tailed unpaired t-test)



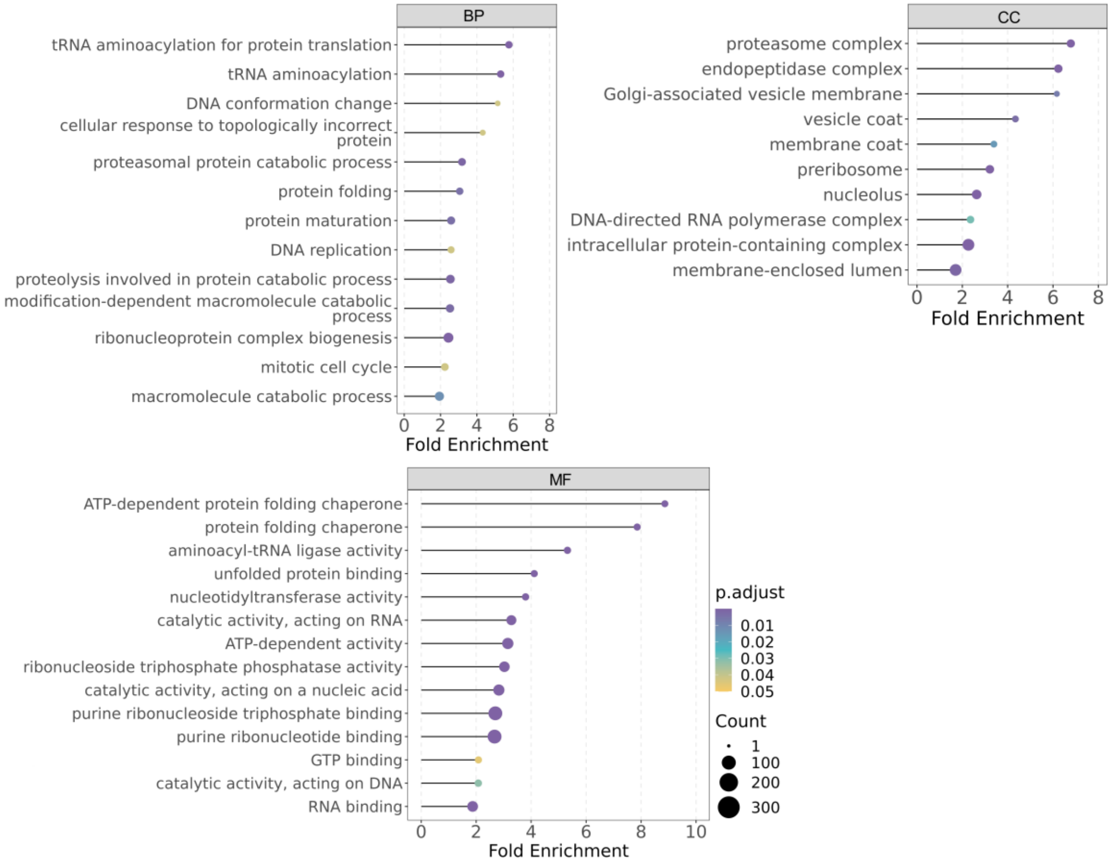
Supplementary Fig. 3- 5. Enriched GO biological process terms (adjusted p-value < 0.05) with the respective FS obtained from our CRISPR-Cas9 screen for identified essential genes associated to each term[1–3]. The FS values for each GO term were found to be significantly lower than those of all genes by unpaired t-test ($p < 0.0001$).



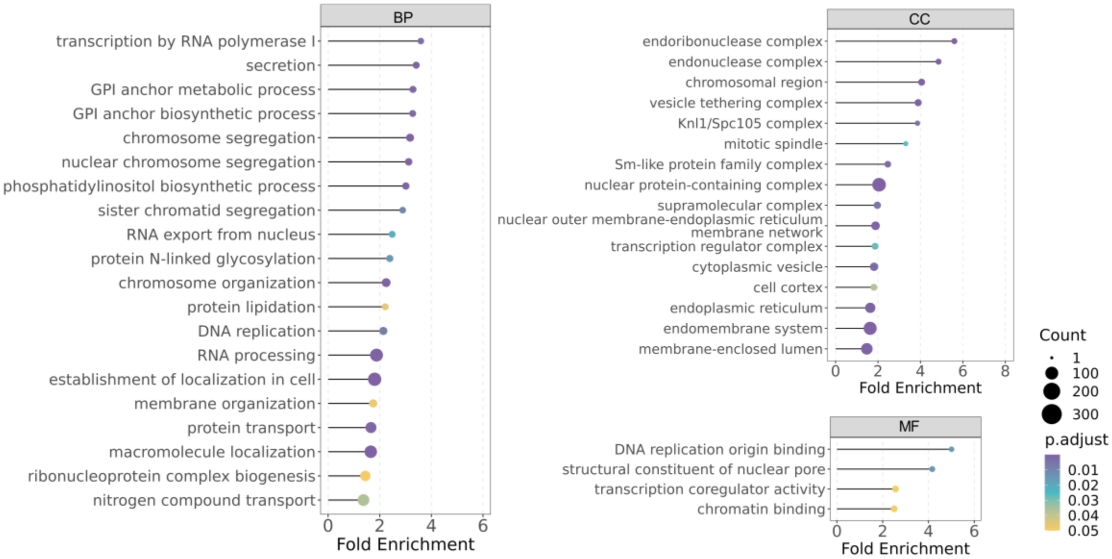
Supplementary Fig. 3- 6. Phylogenetic tree of the analyzed yeast species.



Supplementary Fig. 3- 7. GO enrichment analysis for the consensus set of essential genes. Over-representation analyses were conducted by clusterProfiler, and only the GO terms highly enriched (adjusted p-value < 0.05) in each category (BP: biological process, MF: molecular function, and CC: cellular component) were presented[1–3]. The count represents the number of genes annotated in a specific term, and fold enrichment is defined as the ratio of the frequency of genes belonging to a specific enriched term in the gene set to the frequency of genes belonging to that term in the genome.



Supplementary Fig. 3- 8. GO enrichment analysis for the yeast core set of essential genes. Over-representation analyses were conducted by clusterProfiler, and only the GO terms highly enriched (adjusted p-value < 0.05) in each category (BP: biological process, MF: molecular function, and CC: cellular component) were presented[1–3]. The count represents the number of genes annotated in a specific term, and fold enrichment is defined as the ratio of the frequency of genes belonging to a specific enriched term in the gene set to the frequency of genes belonging to that term in the genome.



Supplementary Fig. 3- 9. GO enrichment analysis for the K. phaffii specific set of essential genes. Over-representation analyses were conducted by clusterProfiler, and only the GO terms highly enriched (adjusted p-value < 0.05) in each category (BP: biological process, MF: molecular function, and CC: cellular component) were presented[1–3]. The count represents the number of genes annotated in a specific term, and fold enrichment is defined as the ratio of the frequency of genes belonging to a specific enriched term in the gene set to the frequency of genes belonging to that term in the genome.

3.10.2 Supplementary tables

Supplementary Table 3- 1. Fold-coverage of the designed library for K. phaffii GS115[4]. More than 98% of the genes in the genome are designed to be targeted by six sgRNAs. Only seventeen genes lacked any designed sgRNAs due to redundancy in their sequence which led to the absence of unique guides.

sgRNA fold-coverage	No. of genes with fold-coverage	% of genes with fold-coverage
6	5132	94.58
5	37	0.68
4	71	1.31
3	70	1.29
2	64	1.18
1	35	0.64
0	17	0.31

Supplementary Table 3- 2. Strains used in this study.

Strain	Genotype	Reference
E. coli TOP10	F- mcrA Δ(mrr-hsdRMS-mcrBC) φ80lacZΔM15 ΔlacX74 recA1 araD139 Δ(ara-leu) 7697 galU galK rpsL (StrR) endA1 nupG λ	Thermo Fisher Scientific
GS115	*his4*	Invitrogen
GS115 *his4*::*CAS9*	*his4*::*CAS9*	This study
GS115 *ku70*	*his4, ku70*	This study
GS115 *his4*::*CAS9 ku70*	*his4*::*CAS9, ku70*	This study

Supplementary Table 3- 3. Plasmids used in this study

Plasmid	Description	Reference
D-227	Contains Cas9 and sgRNA expression cassettes in *K. phaffii*	[5]
pCRISPRpp	sgRNA library backbone, contains PpHIS4 gene, PARS1 and AmpR	This study
pCRISPRyl	CRISPR/Cas9 vector for Yarrowia lipolytica, with AvrII site for sgRNA insertion	Addgene #70007
BB3cK_pGAP_23*_pTEF_Cas9	hCas9 under control of Tef1 for direct cloning of HH-sgRNA-HDV PCR products and episomal expression in P. pastoris and G418 selection	Addgene #104909 [6]
pMJA089	Expresses human Lysozyme in Pichia pastoris	Addgene #128518[7]

Supplementary Table 3- 4. Primers used in this study.

Primer name	Primer sequence	Use
UDA.F	ATCAACGCTGTTCAACAAAATCTCAACA	*HIS4* loci upstream donor arm amplification
UDA.R	TGCACATTAACTTGAAGCTCAGTCG	*HIS4* loci upstream donor arm amplification
DDA.F	AGGACCAATTTGAGGAGCTGAT	*HIS4* loci downstream donor arm amplification
DDA.R	TGTTCCTTGGTGTATCCTGGCT	*HIS4* loci downstream donor arm amplification
D227_UDA.F	GAGCTTCAAGTTAATGTGCAATTAAGTAATAGCAAGGTAAAGTAATACAGGGAGT	D-227 vector linearization for upstream donor arm integration
D227_UDA.R	GATTTTGTTGAACAGCGTTGATGACTCCCAAGTCTAAGGACTTGA	D-227 vector linearization for upstream donor arm integration
D227_DDA.F	GCCAGGATACACCAAGGAACAACCATGGGATATGTTTCACGTTTTGT	D-227 vector linearization for downstream donor arm integration
D227_DDA.R	CAGCTCCTCAAATTGGTCCTATGAAAGAGTGAGAGGAAAGTACCTG	D-227 vector linearization for downstream donor arm integration
HIS4_ID.F	CTAAACGAAAGACTACATTTCTAGATGAGTTTGCC	PCR amplification of the *HIS4* loci to check for Cas9 integration
HIS4_ID.R	CCTGACGTTATCTATAGAGAGATCAATGGCTC	PCR amplification of the *HIS4* loci to check for Cas9 integration
Ori_AmpR.F	AATACGGTTATCCACAGAATCAGGGGA	Ori_AmpR PCR amplification from pCRISPRyl
Ori_AmpR.R	GGAAATGTGtGCGGAACC	Ori_AmpR PCR amplification from pCRISPRyl
PARS1.F	ACGAGGCCCAGATCCTCTATTAATTAACCTAGGGGTACCTTCAAGTTTCGTTAAGCAGGA	PARS1 PCR amplification from the genome
PARS1.R	CCTGATTCTGTGGATAACCGTATTACTAGTGATTGATATTGGAACCTGCTGTCATT	PARS1 PCR amplification from the genome
pTEF1.F	AATAGGGGTTCCGCGCACATTTCCGCATGCgcatcaccatctgaatatttgaccgct	TEF1 promoter amplification from BB3cK_pGAP_23*_pTEF_

		Cas9 plasmid
pTEF1.R	gcaggtagcaagggaaatgtcatGGTTACCgaccgcccttagattagattgctatgc	TEF1 promoter amplification from BB3cK_pGAP_23*_pTEF_Cas9 plasmid
CYC1t.F	gaaactgggacttatttaaGGGCCCtcatgtaattagttatgtcacgcttac	CYC1 terminator amplification from BB3cK_pGAP_23*_pTEF_Cas9 plasmid
CYC1t.R	ATTCAAGCTAATATGGCTGATGATCCTCTAACCTACTACGcatgaattagcgccagcttg	CYC1 terminator amplification from BB3cK_pGAP_23*_pTEF_Cas9 plasmid
ptRNA1_tRNA1.F	AGCCATATTAGCTTGAATGATTGGATTTTTTGTAGCTTTATAAGCAGCTTTTTCTTGAAG	gRNA expression cassette amplification from D-227
ptRNA1_tRNA1.R	GGTTAATGTCATGATAATAATGGTTTCTTAgacgtAAAAAAAAGCACCGACTCGGTGCC	gRNA expression cassette amplification from D-227
PpHIS4.F	atgacatttcccttgctacctg	Pichia pastoris HIS4 gene from pIB1
PpHIS4.R	aagcgtgacataactaattacatgaGGGCCCttaaataagtcccagtttctccatacg	Pichia pastoris HIS4 gene from pIB1
5'_60mer	tagtggtagaaccaccgcttgtcgcgcggtagaccggggttcaattccccgtcgcggagc	5' 60mer linker added to the designed sgRNAs in the library
3'_60mer	gttttagagctagaaatagcaagttaaaataaggctagtccgttatcaacttgaaaaagt	3' 60mer linker added to the designed sgRNAs in the library
qPCR_GW.F	GCGCCTTATCCGGTAACTATC	qPCR experiment primer to count the copy number of miniprepped library from GS115 cells
qPCR_GW.R	CTACATACCTCGCTCTGCTAATC	qPCR experiment primer to count the copy number of miniprepped library from GS115 cells
NGS1.F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGtgtagaccggggttcaattccc	Forward Illumina primer used for NGS
NGS2.F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGAT	Forward Illumina primer used for NGS

	CTGTagtgtagaccggggttcaattccc	
NGS3.F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTCAGTagtgtagaccggggttcaattccc	Forward Illumina primer used for NGS
NGS4.F	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTTCCAGTagtgtagaccggggttcaattccc	Forward Illumina primer used for NGS
NGS1.R	CAAGCAGAAGACGGCATACGAGATTCGCCTTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTGATAACGGACTAGCCT	Reverse Illumina primer used for NGS
NGS2.R	CAAGCAGAAGACGGCATACGAGATATAGCGTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTGATAACGGACTAGCCT	Reverse Illumina primer used for NGS
NGS3.R	CAAGCAGAAGACGGCATACGAGATGAAGAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTGATAACGGACTAGCCT	Reverse Illumina primer used for NGS
NGS4.R	CAAGCAGAAGACGGCATACGAGATATTCTAGGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTGATAACGGACTAGCCT	Reverse Illumina primer used for NGS
NGS5.R	CAAGCAGAAGACGGCATACGAGATCGTTACCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTGATAACGGACTAGCCT	Reverse Illumina primer used for NGS
NGS6.R	CAAGCAGAAGACGGCATACGAGATGTCTGATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTGATAACGGACTAGCCT	Reverse Illumina primer used for NGS
NGS7.R	CAAGCAGAAGACGGCATACGAGATTTACGCACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTGATAACGGACTAGCCT	Reverse Illumina primer used for NGS
NGS8.R	CAAGCAGAAGACGGCATACGAGATTTGAATAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTGATAACGGACTAGCCT	Reverse Illumina primer used for NGS
NGS9.R	CAAGCAGAAGACGGCATACGAGATTGCTGAGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAAGTTGATAACGGACTAGCCT	Reverse Illumina primer used for NGS

Supplementary Table 3- 5. Primers used in this study.

gRNA sequence	Targeting gene
AGTTGAGTCATCCGTTACAG	*HIS4*
AAGCAATACGACATCCACGA	*KU70*
GAAGCTCAAGGAATCAAGAA	GS115_000439_6.0
AAGCACTCAGACTCTCCCTT	GS115_003019_6.0
AAACATCATTATCGGGCGTG	GS115_000695_1.0
ATGTCGCTCTTCTTCATAGA	GS115_001683_4.0
TGGAGGTCTAATCATAACTG	GS115_003470_2.0
CGGAGAGTTACAATTCTCCG	GS115_002498_2.0
GGCAATATGCCATTGCCACC	GS115_001188_2.0
CTTTGTGGCCTCTAAATGAG	GS115_002303_4.0
ACATCCTGCTGCAAAGATGT	GS115_003685_4.0
GTCCACCTCCAAATCCACCA	GS115_005273_3.0
GGACTCAGATCCAGACTCGG	GS115_004564_1.0
CCAAAGGATCCCAGGCGCTA	GS115_000983_5.0
AAGCAATACGACATCCACGA	GS115_003415_5.0
CCAGAGAATCAAATTGCCAA	GS115_004871_6.0
ATATAGGGATCTTCCAAAGG	GS115_002757_4.0
GCAGAAGCGCTCCAGTACGA	GS115_004368_2.0
GAGCTTCAGCAGTGGTACAA	GS115_000502_5.0
GCCACCCAAATCCACCCTCG	GS115_001665_3.0
CTATGAAACAAAAACGTTCT	GS115_001679_6.0
CTCATTCAAAGAATTGAATG	GS115_000619_4.0
AATTCCTTCAACAAATCCGG	GS115_002448_5.0
TCCAATGACATGAAAGCCAT	GS115_003547_2.0
TCACGTCGACTGATGCGCAT	GS115_003476_1.0
GAAGCTCAAGGAATCAAGAA	GS115_000439_6.0
AAGCACTCAGACTCTCCCTT	GS115_003019_6.0
TGGCGAACTAATCTCCGTAC	GS115_000039_3.0
GGTGTCTCTTAGCTTCTCAG	GS115_004484_4.0
TACTGTTTACAAGGGAACGA	GS115_002074_3.0
TGAGATGGCTGAGCGGAGGG	GS115_001690_3.0
GATGAAAATGAGATTCTGGC	GS115_004666_5.0
TACAGCAGCCATCTTGTAAG	GS115_002712_5.0
CATAGTGAATCAGTCAACCT	GS115_005262_1.0
GTGCTTTCAACATAATCTGG	GS115_002014_1.0
GGTGAAACAAGAGGTTGCGA	GS115_000024_2.0
ATCTGGTAGTGGAGAACCGG	GS115_003079_4.0
ACTTGCTGAACAAATCGGGT	GS115_002850_6.0
GCTTCTATACAAATCGGATG	GS115_002100_6.0
GTTGAGACTGTCAGAAATTG	GS115_005301_5.0

GCTCAGGACAGATAACGTGC	GS115_001825_5.0
GTTCACAGAGGCCTACAGAC	GS115_000033_6.0
AAGAAACATTAACACAACAT	GS115_004475_2.0
GTGAGATTTGAGATTCAAGG	GS115_001413_2.0
ACCAACGGTAAAGTTCCTGA	GS115_003572_4.0
GTGGCCAAGAGAACTCCAGC	GS115_001564_4.0
GTCCTCACGATCGTTGACAA	GS115_001871_2.0
TGGTGTAACAAATATCGCCT	GS115_002147_5.0
TGCCAAACGAAGAATCAACC	GS115_001537_3.0
TGCAACCATTGAGTAGTAGT	GS115_002312_4.0
GGAGCTCCAGTGGGGACAGT	GS115_001830_4.0
ACTATGACTCGGAAGAAAGA	GS115_001234_2.0
AGACACACAGGATAGCGGGC	GS115_003613_4.0
TCTACGACGCCGTTTAGCAC	GS115_004444_4.0
ACCGAAGAAGATGATGCGGA	GS115_000471_2.0
GTAAAAACCCTATCAGGCCG	GS115_001778_2.0
TCCAACTCCAAGGATACCAT	GS115_003545_6.0

Supplementary Table 3- 6. Comprehensive breakdown of quality scores used to design the final library.

Quality score	Criteria
1	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 0, MM3 = 0
2	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 0, MM3 = 1
3	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 0, MM3 = 2
4	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 0, MM3 = 3
5	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 1, MM3 = 0
6	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 1, MM3 = 1
7	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 1, MM3 = 2
8	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 1, MM3 = 3
9	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 2, MM3 = 0
10	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 2, MM3 = 1
11	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 2, MM3 = 2
12	Seed_MM0 = 0, MM0 = 0, Self_complementarity = 0, MM1 = 0, MM2 = 2, MM3 = 3
13	Seed_MM0 = 1, MM0 = 1

Supplementary Table 3- 7. Number of sgRNAs in the designed library with their respective quality score.

Quality Score	No. of gRNAs	% of the library
1	31433	99.36
2	61	0.19
3	22	0.07
4	5	0.02
5	18	0.06
6	6	0.02
7	1	0.003
8	1	0.003
9	0	0
10	0	0
11	2	0.006
12	1	0.003
13	84	0.26
Non-targeting	350	1.089155
Total	32485	100

Supplementary Table 3- 8. Transformation efficiencies measured as $\times 10^6$ transformants for 10 pooled transformations, for all replicates in the control and treatment strains.

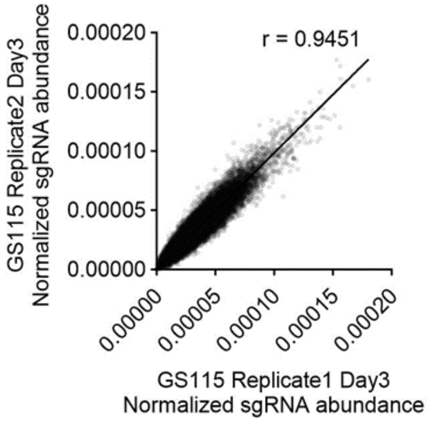
	Transformation efficiency ($\times 10^6$)		
Strain	Replicate 1	Replicate 2	Replicate 3
GS115	7.03	10.93	6.3
GS115 Cas9	8.35	6.85	10.7
GS115 Δku70	5.1	5.4	8.86
GS115 Δku70 Cas9	4.9	7.55	9.55

Supplementary Table 3- 9. Library plasmid copy number and fold-coverage from miniprepped samples for all strains and replicates.

Strain-Replicate-Day	Plasmid copy number	Fold-coverage
GS115-Rep1-Day3	3.21E+08	9441
GS115-Rep1-Day6	4.40E+08	12941
GS115-Rep2-Day3	1.36E+09	40000
GS115-Rep2-Day6	1.98E+08	5824
GS115-Rep3-Day3	1.32E+09	38824
GS115-Rep3-Day6	1.04E+08	3059
GS115 Cas9-Rep1-Day3	4.08E+08	12000
GS115 Cas9-Rep1-Day6	1.50E+08	4412
GS115 Cas9-Rep2-Day3	1.10E+09	32353
GS115 Cas9-Rep2-Day6	1.93E+08	5676
GS115 Cas9-Rep3-Day3	6.62E+08	19471
GS115 Cas9-Rep3-Day6	1.43E+08	4206
GS115 Δku70-Rep1-Day3	5.72E+07	1683
GS115 Δku70-Rep1-Day6	1.40E+08	4129
GS115 Δku70-Rep2-Day3	1.63E+08	4803
GS115 Δku70-Rep2-Day6	5.28E+07	1552
GS115 Δku70-Rep3-Day3	1.74E+08	5115
GS115 Δku70-Rep3-Day6	3.72E+07	1095
GS115 Δku70 Cas9-Rep1-Day3	1.89E+08	5562
GS115 Δku70 Cas9-Rep1-Day6	8.90E+07	2616
GS115 Δku70 Cas9-Rep2-Day3	2.69E+08	7918
GS115 Δku70 Cas9-Rep2-Day6	1.69E+08	4974
GS115 Δku70 Cas9-Rep3-Day3	1.19E+08	3497
GS115 Δku70 Cas9-Rep3-Day6	1.48E+08	4344

Supplementary Table 3- 10. Correlation of normalized sgRNA abundance between biological replicates. The plot shows the normalized sgRNA abundance between replicate 1 and 2 for GS115. Linear regression between replicate 1 and 2 yields a Pearson coefficient of 0.9451. The table represents the correlation of normalized sgRNA abundance for all strains and all possible combinations of replicates.

Strain	Time	Replicate	Pearson r
GS115	Day3	1 vs. 2	0.89
		1 vs. 3	0.8874
		2 vs. 3	0.8911
GS115	Day6	1 vs. 2	0.8236
		1 vs. 3	0.8216
		2 vs. 3	0.8453
GS115 Cas9	Day3	1 vs. 2	0.9829
		1 vs. 3	0.9852
		2 vs. 3	0.9823
GS115 Cas9	Day6	1 vs. 2	0.9792
		1 vs. 3	0.98
		2 vs. 3	0.9821
GS115 Δku70	Day3	1 vs. 2	0.8759
		1 vs. 3	0.8859
		2 vs. 3	0.8962
GS115 Δku70	Day6	1 vs. 2	0.6654
		1 vs. 3	0.6936
		2 vs. 3	0.8015
GS115 Δku70 Cas9	Day3	1 vs. 2	0.9799
		1 vs. 3	0.9836
		2 vs. 3	0.9844
GS115 Δku70 Cas9	Day6	1 vs. 2	0.9522
		1 vs. 3	0.9745
		2 vs. 3	0.9531



References

1. Carlson, M. & Pages, H. AnnotationForge: tools for building SQLite-based annotation data packages. R Packag. version 1.32. 0. Preprint at (2020).
2. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).
3. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
4. Labun, K. *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* **47**, W171–W174 (2019).
5. Dalvie, N. C. *et al.* Host-Informed Expression of CRISPR Guide RNA for Genomic Engineering in Komagataella phaffii. *ACS Synth. Biol.* **9**, 26–35 (2020).
6. Gassler, T., Heistinger, L., Mattanovich, D., Gasser, B. & Prielhofer, R. CRISPR/Cas9-Mediated Homology-Directed Genome Editing in Pichia pastoris. *Methods Mol. Biol.* **1923**, 211–225 (2019).
7. Yang, J. *et al.* Hygromycin-resistance vectors for gene expression in Pichia pastoris. *Yeast* **31**, 115–125 (2014).

Chapter 4. Ribo-seq guided design of enhanced protein secretion in *Komagataella phaffii*

4.1 Abstract

The production of recombinant proteins requires the precise coordination of various biological processes, including protein synthesis, folding, trafficking, and secretion. Production of native proteome can impose bottlenecks on these complex networks that are not optimized for overproduction of a target protein. The methylotrophic yeast *Komagataella phaffii* is well-known and used for its high capacity to produce recombinant proteins. To investigate these bottlenecks in this industrially relevant yeast, we employed ribosome profiling to quantify the global and early secretory demands before and after methanol induction in a strain engineered to produce and secrete human serum albumin (HSA). Through next-generation sequencing and ribosome profiling, we identified key host proteins that constrain biogenetic machinery in *K. phaffii* during heterologous protein expression. By targeting these proteins using the CRISPR-Cas9 system, we achieved a 35% increase in HSA secretion, demonstrating a rational approach to optimizing secretion and overcoming bioproduction bottlenecks. This strategy offers valuable insights into the metabolic and secretory demands of *K. phaffii* and highlights new targets for strain engineering.

This chapter under preparation as an article. The original citation is as follows: Tafrishi, A., Alva, T., Chartron, J., & Wheeldon, I. (2024). Ribo-seq guided design of enhanced protein secretion in *Komagataella phaffii*

4.2 Introduction

Engineering structurally and functionally diverse biologics such as enzymes, materials, and therapeutics is an essential task in biotechnology and the biopharmaceutical industry [1]. Biologic therapeutics represent the area of highest growth in the medical industry [2] with a valued global market size of USD 511.04 billion in 2024 and a compound annual growth rate (CAGR) of 10.4% from 2024 to 2033 [3]. *Komagataella phaffii* (*K. phaffii*), commonly known as the 'biotech yeast'[4], stands out among the fungal kingdom for its ability to grow to high cell densities as it prefers respiration over fermentation. It secretes high levels of recombinant proteins while secreting low levels of endogenous proteins which results in lower costs and time of downstream protein purification. *K. phaffii* is a methylotrophic yeast and is able to metabolize methanol as its primary carbon source using tightly regulated alcohol oxidase genes (*AOX1* and *AOX2*) with the extremely strong *AOX1* promoter being widely employed for controllable expression of recombinant proteins [4].

Industrial bioproduction in *K. phaffii* typically involves growing cells in glycerol-based media before switching to methanol-based media for heterologous induction, using the *AOX1* promoter for precise protein expression [5–7]. The yields of heterologous protein products in yeasts often suffer from bottlenecks in biogenesis [8,9]. Various strategies have been employed to increase heterologous production including optimization of growth conditions (such as methanol concentration in *K. phaffii*), modifying mRNA structural elements, engineering signal sequences, and modification of genes involved in the secretory pathway [10–18]. While these approaches can improve secretion, the resulting increases in production titers are often incremental, and optimizations that are effective in one context may not translate well to others [19,20].

Protein secretion is complex and prone to bottlenecks, particularly during protein trafficking through the endoplasmic reticulum (ER), the rate-limiting step in production [21]. Heterologous trafficking is contingent on the recognition and binding of N-terminus hydrophobic motifs signal sequences by a signal recognition particle (SRP) [21,22]. SRP guides the ribosome nascent chain (RNC) complex to the ER membrane where they associate with translocons by interaction of SRP's cognate receptor and translocate co-translationally [23]. Access to protein folding chaperones in *K. phaffii* is made more difficult as previous studies show that an equal amount of nascent polypeptides translocate across the ER co-translationally and post-translationally[24]. This is additionally problematic as the heptameric post-translational Sec-translocon requires the same subunits as the hexameric Sec-translocon as well as an additional subunit. Misfolded proteins in the ER are not transported to the Golgi and instead activate the unfolded protein response (UPR). Proteins that activate the UPR are often destroyed using the ER-associated degradation pathway (ERAD) [25].

ribosome profiling (Ribo-seq) is a high-throughput sequencing technique that measures protein synthesis by assessing ribosome abundance at each codon in a transcript [26]. Compared to RNA-seq, Ribo-seq offers a closer correlation with standard proteomics and is much higher throughput than mass spectrometry, while still accurately predicting mature protein stoichiometry [27–29]. However, Ribo-seq comes with challenges, particularly in the isolation of ribosome-protected mRNA footprints, which requires effective rRNA subtraction methods [30–33].

To enhance secretion in *K. phaffii*, we used next-generation sequencing to analyze the translatome under heterologous conditions for rational strain engineering. We explored methanol metabolism and employed Ribo-seq alongside ER trafficking predictions to generate data reflecting translatome variations in wildtype and human serum albumin (HSA)-expressing strains. HSA is a ~67 kDa protein with semi-complex folding requirements and minimal glycosylation. We introduce a novel rRNA subtraction technique that employs commercially available depletion agents and probe-directed degradation with DSN, demonstrating superior performance compared to other strategies used in yeast studies [33–36]. By modeling metabolic and secretory demands, we identified key host cell proteins involved in early secretory pathways and used CRISPR-Cas9 to knockout these genes, resulting in a 35% increase in HSA secretion. This approach reveals new targets for optimizing secretion. Our methodology reveals novel insights into these conditions and allows for a rational approach to widen secretion bottlenecks by providing new targets for modification that would not have otherwise been predicted.

4.3 Results

4.3.1 Surveying translation with Ribo-seq

We used the high throughput technique Ribo-seq to measure protein synthesis for GS115 Mut$^+$ and GS115 Mut^S ALB cultures collected before, and at 3 and 24 hours after, methanol induction (**Fig. 4-1a**). Ribo-seq uses a non-specific nuclease to break down nucleic acids, including mRNA, that aren't protected by ribosomes. To sequence ribosome protected mRNA fragments and reveal translational dynamics, ribosome derived RNA (rRNA) first needs to be depleted. We found that previous strategies to remove rRNA contamination in *K. phaffii* collected at log-phase growth in YPD media [13] were not sufficient for generating high quality Ribo-seq libraries where cells are collected at different growth stages and in different media. Our datasets agreed with previous Ribo-seq analyses [18] and revealed that a subset of rRNA represented the majority of rRNA contamination (**Supplementary Table 4-1**). Complementary oligos from pre-induction samples were used for probe-directed DSN treatment, reducing rRNA contamination from 88% to 10% in the pre-induction sample, 87% to 20% at 3 hours post-induction, and 93% to 62% at 24 hours post-induction. Before and three hours after induction, nearly all reads mapped to open reading frames (ORFs) as only 2% of reads mapped to untranslated regions (UTRs). Twenty-four hours after induction, however, we observed increased reads mapped outside of annotated ORFs as nearly 7% of reads mapped to UTRs. This was particularly true for genes like *GLN1* and *GCN4* that have previously been shown to have increased read density at 5'UTRs in response to stress [18](**Fig. 4-b**).

Our data revealed genome-wide coverage of expression, as up to 93% and 94% of *K. phaffii*'s 5,330 annotated protein-encoding genes were detected on average for GS115 Mut^{+} and GS115 Mut^{S} ALB, respectively (**Supplementary Fig. 4-1** and **Supplementary Data 1 and 2**). Before making intra- and inter-sample comparisons of expression levels, we first sought to normalize reads (**Fig. 4-1c**). First, footprint sized fragments were used to generate computational masks for codons with a propensity to map to multiple locations of the genome. Next, reads per codon for the first 500 codons were normalized in all genes to account for positional counting biases in codons that were masked. Finally, we determined gene read count thresholds for comparing expressions between samples. To calculate these thresholds, we used biological replicates in the GS115 Mut^{S} ALB strain. In doing so, genes were binned according to the probabilistic distribution of the summed read counts per gene between each replicate. Binned genes' read counts were normalized by the total read counts between both replicates. The standard deviation of each gene's normalized reads with respect to their bin's read count value was used to calculate read count thresholds necessary to reduce inter-replicate variability. These thresholds were used to predict read count thresholds for all samples (including those without replicates) as a function of their summed reads. This conservatively calculated read count thresholds between 52 reads to 573 reads, where samples with greater total reads had greater count thresholds. These criteria filtered approximately 1% of total nascent chains calculated per sample.

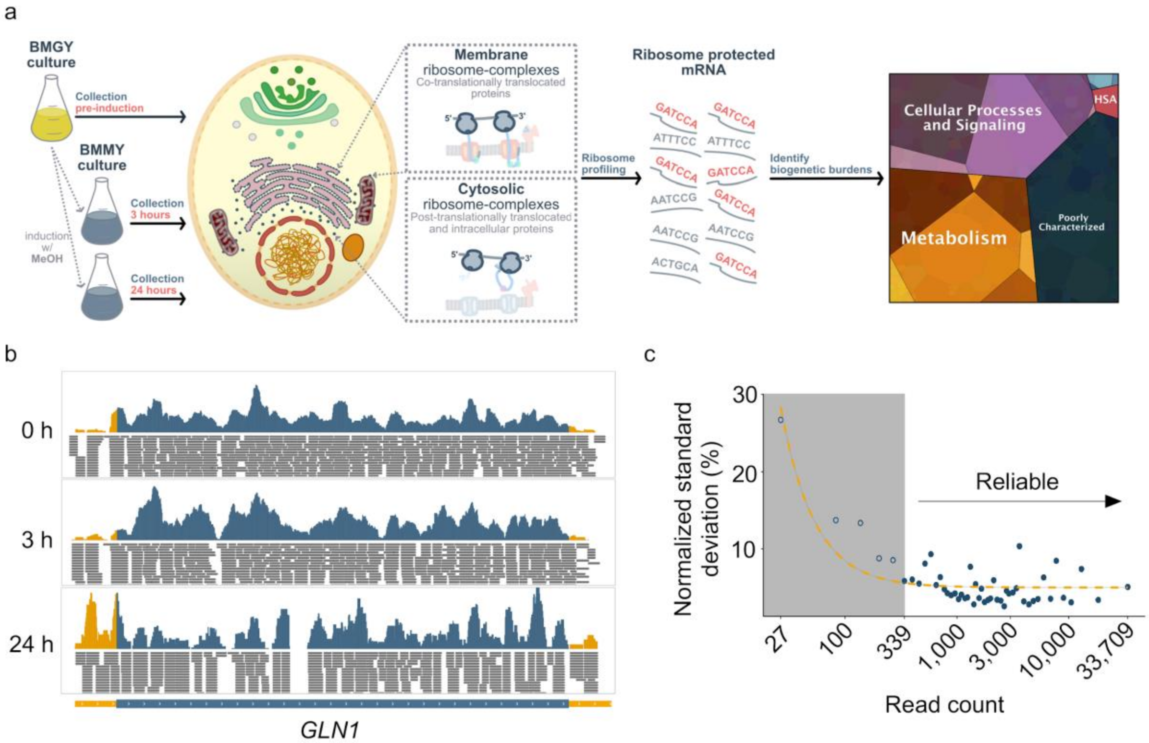


Fig. 4- 1. a) Overview of heterologous expression and Ribo-seq. Starter cultures were grown in buffered glycerol media (BMGY), with one-third collected and flash-frozen. The remaining culture was transferred to buffered methanol media (BMMY) for heterologous induction. Samples were collected 3 and 24 hours after induction. Membrane-associated and cytosolic ribosomes were isolated, corresponding to co- and post-translationally translocated proteins, respectively. mRNA footprints were then isolated from ribosomes before Illumina sequencing. **b)** Ribosome abundance on transcripts under heterologous conditions for *GLN1*. The represented transcript reads are from GS115 Mut^S ALB cultures collected from the pre-induction sample cultured in BMGY media, 3 h and 24 h after induction in BMMY media. The bottom blue band shows the ORF while the yellow band shows the UTRs. After 24 hours in the induction media a much higher proportion of reads map to the 5'UTR. Images are modified screen captures from Integrated Genome Viewer (MIT). **c)** Determining reads per gene thresholds. Biological replicates were used to determine read count thresholds when comparing genetic expression between data sets. Total read counts per gene were calculated by summing reads per gene for each replicate. Genes were binned according to this total read count value. Replicate read fractions were calculated by dividing read counts per gene by their bin value. Standard deviations of replicate read fractions were computed across each bin. Standard deviations were fit to replicate reads per gene using a generalized exponential decay model. Minimum read thresholds were calculated as the inflection point in this regressed curve. When reads per gene are fewer than this threshold, counting errors predominate inter-replicate variability. When reads per gene exceed this threshold, other sources of error predominate.

4.3.2 Translational landscape under heterologous conditions

We used Ribo-seq to survey nascent chain production in GS115 Mut$^+$ and GS115 Mut^S ALB cultures collected before methanol induction and 3 and 24 hours after methanol induction (**Supplementary Data 1** and **2**). **Fig. 4-2a** shows gene expression fold changes 3- and 24-hours after methanol induction. The majority of the genes, 63% in GS115 Mut$^+$ and 83% in GS115 Mut^S ALB, maintain the same expression levels three hours after methanol induction. However, after 24 hours in the heterologous conditions, the number of genes maintaining the same expression levels decreases to 22% and 32% in GS115 Mut$^+$ and GS115 Mut^S ALB.

Next, we studied the total nascent chain production related to specific ontological categories including cellular processes and signaling, information storage and processing, and metabolism pathways as well as genes with functions that have yet to be characterized in each strain (**Fig. 4-2b**). The production of nascent chains shifts under heterologous conditions, becoming more pronounced after 24 hours. The cumulative nascent chain production for genes involved in cell processes and signaling is relatively conserved over time and is an indication of these functions' vitality. However, some genes belonging to this category that are involved in UPR like *HAC1* show 6.3- and 3.4-fold increased expression after 24 hours of methanol induction in GS115 Mut$^+$ and GS115 Mut^S ALB, respectively.

In contrast, the production of nascent chains related to information storage and processing decreased significantly by 56% in GS115 Mut$^+$ and 49% in GS115 Mut^S ALB. This decrease was primarily driven by the reduced expression of genes involved in translation and ribosome biogenesis, which dropped by 73% in GS115 Mut$^+$ and 57% in GS115 Mut^S ALB (**Fig. 4-3**, refer to **Supplementary Fig. 4-2** for fold change results after 3 hours in the induction media). This decrease was accompanied by increased expression of transcription- as well as replication, recombination and repair-related genes.

Uncharacterized genes showed increased expressions of 86% and 129% in GS115 Mut$^+$ and GS115 Mut^S ALB strains after 24 hours of induction. The most differentially expressed uncharacterized proteins are those predicted to localize in the peroxisome, where energy is produced in the methanol utilization (MUT) pathway with a 4.2- and 3.4-fold increased expression in GS115 Mut$^+$ and GS115 Mut^S ALB.

Nascent chain production for metabolism-related genes increased by 81% and 45% in GS115 Mut$^+$ and GS115 Mut^S ALB strains, respectively. Genes involved in the synthesis, transport, and catabolism of secondary metabolites, lipids, and carbohydrates are affected the most. Although amino acid transport and biosynthesis-related genes are not the most differentially expressed overall, we observe significant differential expression in specific genes, such as *GLN1*, which shows a 7.1-fold increase in GS115 Mut$^+$ and 3.23-fold increase in GS115 Mut^S ALB, and *PUT1*, with a 670-fold and 6.1-fold increase in GS115 Mut$^+$ and GS115 Mut^S ALB.

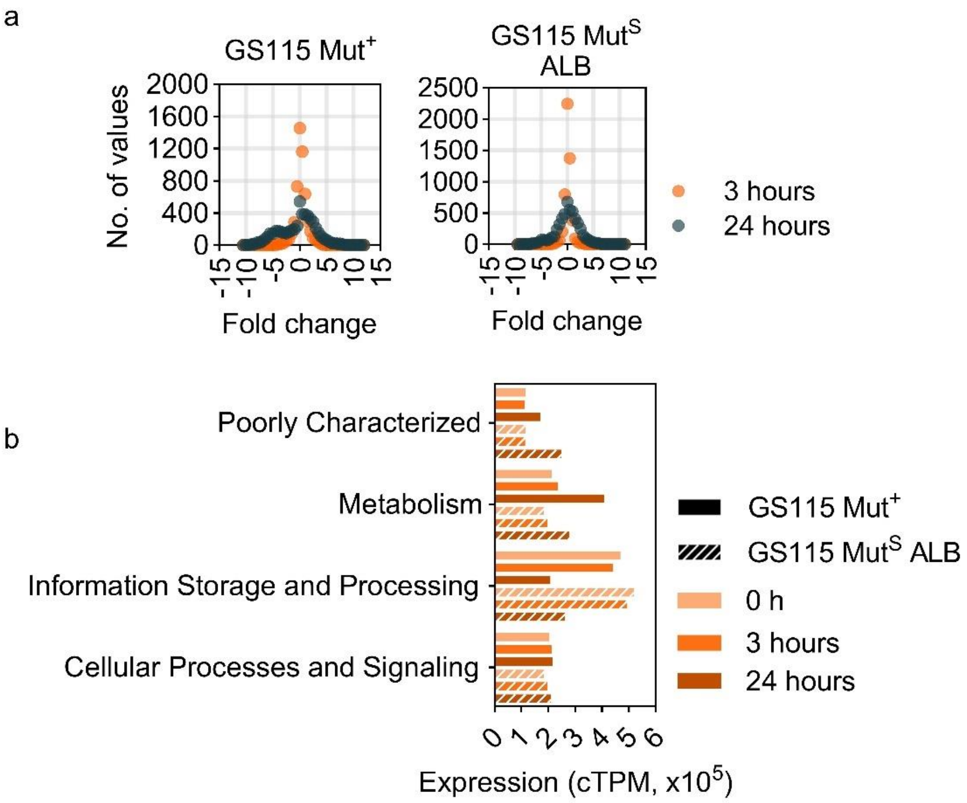


Fig. 4- 2. *G*ene expression changes post induction *for GS115 Mut*$^+$ *and GS115 Mut*^{S} *ALB strains.* **a)** Gene expression fold change distribution 3 and 24 hours after methanol induction for GS115 Mut$^+$ and GS115 Mut^S ALB strains. Most of the genes display constant levels of expression 3-hour post-induction, whereas gene expression levels dramatically change 24 hours after methanol induction for both strains. **b)** *Total nascent chain production* of genes belonging to cellular processes and signaling, metabolism, information storage and processing, and poorly characterized categories. Genes involved in information storage and processing experience reduced expression whereas genes in the cellular processes and signaling pathways retain their expression levels.

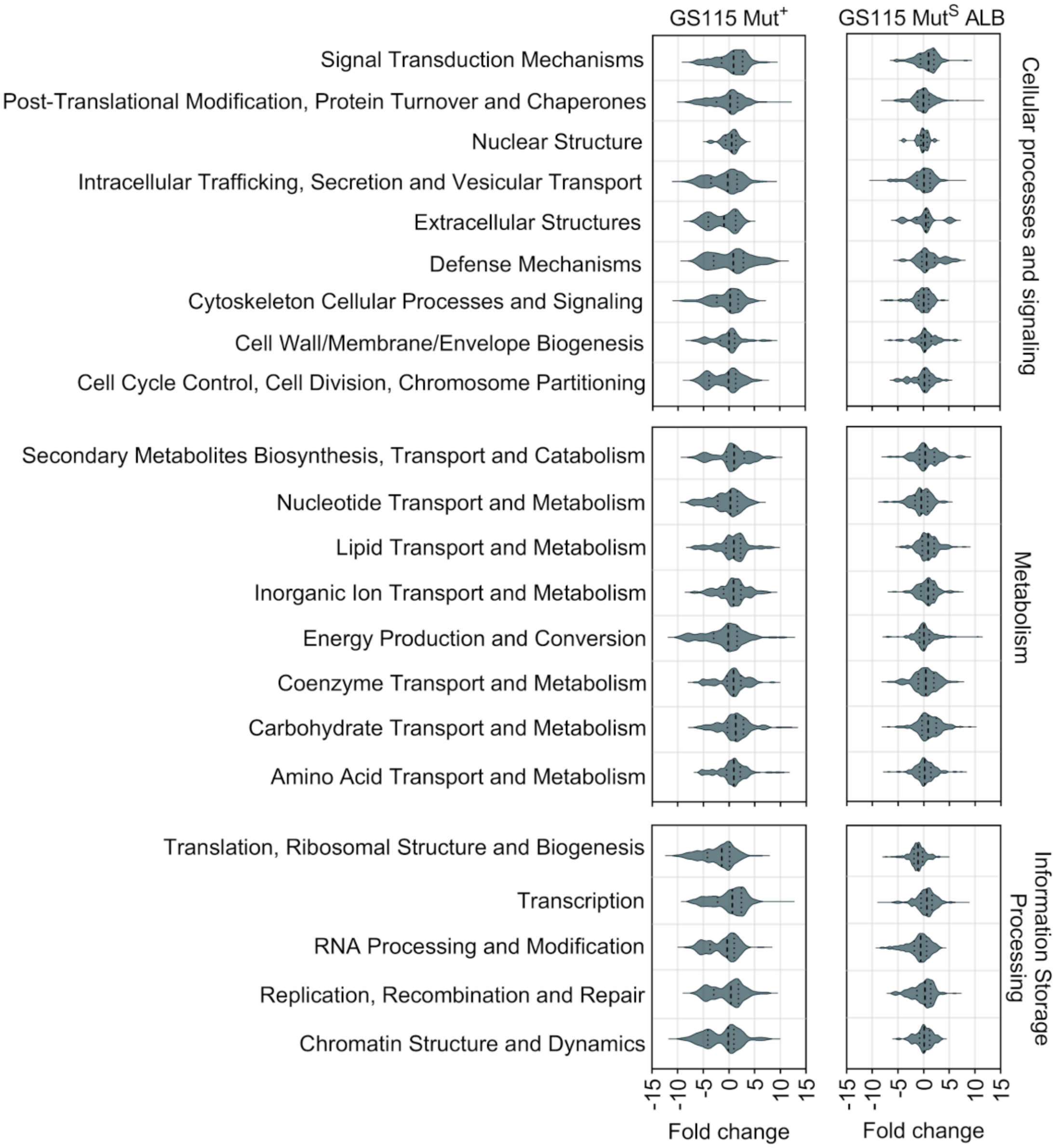


Fig. 4- 3. Fold change production of nascent chains belonging to different ontological categories. Fold change is compared in GS115 Mut$^+$ and GS115 Mut^S ALB 24-hours after induction in buffered methanol media (BMMY). Fold change is calculated as the log2 ratio of expressed genes 24 hours after methanol induction compared to the expression levels before induction. Cells struggle to metabolize methanol after induction and nearly all genes involved in metabolism are positively differentially expressed. As this occurs, expression of genes involved in translation and ribosome biogenesis concomitantly decreased.

4.3.3 Translational changes in the methanol utilization and stress response pathways under heterologous conditions

Gene expression levels in the methanol utilization pathway shift under heterologous conditions (**Fig. 4-4**). The MUT pathway is divided into an assimilative branch, which produces biomass from formaldehyde, and a dissimilative branch, which generates CO_2 from methanol and NADH for ATP production[19]. As part of the assimilative branch of the MUT, genes involved in the pentose phosphate pathway (PPP) also experience differential expression levels. Additionally, the reactive oxygen species (ROS) defense mechanisms are upregulated in the presence of methanol, as the oxidation of methanol to formaldehyde produces reactive hydrogen peroxide.

We find a 58- and 19-fold increased expression of peroxisomal encoding genes 24 hours after growth in methanol media in GS115 Mut^+ and GS115 Mut^S ALB, respectively (**Fig. 4-4a**). In the MUT pathway, AOX is produced exclusively in the presence of methanol and the absence of glucose, breaking down methanol into hydrogen peroxide and formaldehyde in the peroxisome. While there are two genes that encode for AOX in GS115 Mut^+, the majority of AOX activity is expressed through *AOX1* as our datasets detect 32-fold greater expression from *AOX1* than *AOX2* after 24 hours of growth in the induction media similar to what is suggested in other studies [20,21]. Although GS115 Mut^S ALB has a deleted *AOX1*, we see a nearly 7-fold increase in the *AOX2* expression levels after 24 hours.

in the assimilative branch, formaldehyde reacts with xylulose 5-phosphate (Xu5P) and produces dihydroxyacetone (DHA) and glyceraldehyde-3-phosphate (GAP) by dihydroxyacetone synthase (*DAS1* (ACN76559.1), *DAS2* (ACN76560.1), and possibly *TKL1*) [22,23]. Our datasets show that translation of *DAS2* occurs more extensively than *DAS1* as it produces 15.8- and 3.7-fold more nascent chains compared to *DAS1* in GS115 Mut^+ and GS115 Mut^S ALB strains after 24 hours. The cumulative expression of *DAS1* and *DAS2* is nearly 7 times higher in GS115 Mut^+ compared to GS115 Mut^S ALB. The genes involved in the PPP related to biosynthesis of Xu5P are on average 46% and 39% overexpressed in GS115 Mut^+ and GS115 Mut^S ALB (**Fig. 4-4b**). Lower expression of genes to convert formaldehyde to other products in the GS115 Mut^S ALB strain reflects the slower methanol assimilation in this strain as a result of the knocked out *AOX1* which results in the slower production of formaldehyde.

Outside of the peroxisome, formaldehyde is dissimilated to formate by formaldehyde dehydrogenase (FLD) and to carbon dioxide by formate dehydrogenase (FDH) for energy production. We see a 16.8- and 10.1-fold increase in expression levels of FLD genes in GS115 Mut^+ and GS115 Mut^S ALB, as well as a 1717.8- and 2179.7-fold increase for *FDH1*. While we see a 7.3 and 3.3-fold increase in summed nascent chain production for all genes involved in the MUT pathway in GS115 Mut^+ and GS115 Mut^S ALB, the greatest increases in expression are for FDH, *AOX1* and *CTA1* in that order.

Within the peroxisome, toxic hydrogen peroxide is decomposed into oxygen and water by catalase (*CTA1*) as the first step of the ROS defense. *CTA1* exhibits differential expression levels of 130.7-fold and 32.7-fold in GS115 Mut^+ and GS115 Mut^S ALB (**Fig. 4-4c**). The lesser *CTA1* expression increase in the HSA-producing strain is attributable to its lack of a functional *AOX1* gene, resulting in a slower methanol assimilation rate and, consequently, reduced hydrogen peroxide production. As hydrogen peroxide causes oxidative stress, we also observed increased expression involved in oxidative stress responses for genes like *YAP1*, 5.3- and 3.5-fold increase in the GS115 Mut$^+$ and GS115 Mut^S ALB strains, and *GSH2* with 12.6- and 1.14-fold increase in the GS115 Mut$^+$ and GS115 Mut^S ALB strains.

Heterologous protein production in GS115 Mut^S ALB also changes the nascent chain production of genes involved in the UPR pathway (**Fig. 4-4d**). After induction, total nascent chain production of genes involved in UPR are decreased for both strains (18% and 30% reduction in the control and ALB-producing strain, respectively) suggesting that the HSA production likely has low impacts on UPR activation. The lower differential changes of many genes related to UPR in the ALB-producing strain including the *IRE1*, *HAC1*, and *KAR2*, the disulfide isomerase *PDI1,* the mitochondrial chaperones *HSP60* and *SSC1,* as well as cytosolic chaperones including *HSP104, CDC48, UBA1, SSA3* involved in disassembling and dragging the misfolded proteins indicate lower activation of the UPR/ERAD pathways in the ALB-producing strain compared to the control strain. This unexpected result has also previously been observed during methanol-induced production of insulin precursor in controlled fed-batch cultures[46-48]. It has been suggested that the observed reduction in the UPR/ERAD pathways are due to other factors not related to the heterologous protein production.

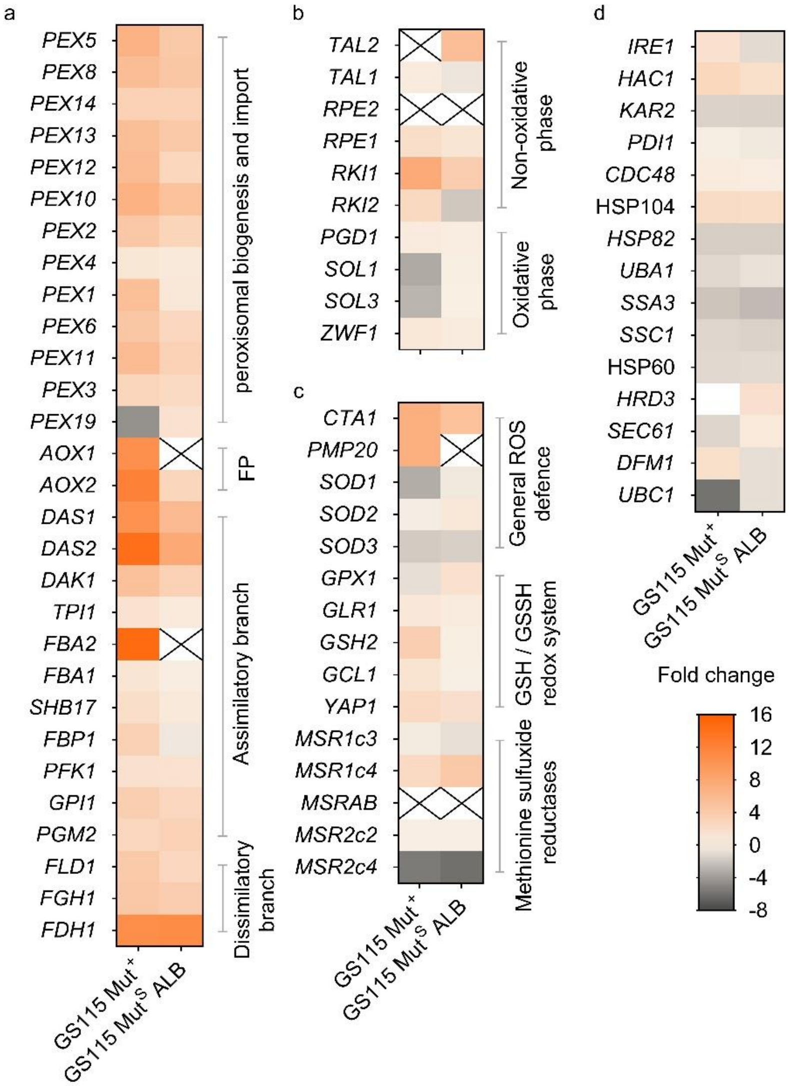


Fig. 4- 4. Regulation of methanol utilization pathway genes. Gene expression fold change for genes involved in **a)** Canonical methanol utilization pathway, **b)** Pentose phosphate pathway, **c)** Reactive oxygen species (ROS) pathway, and **d)** Unfolded protein response (UPR). Fold change is calculated as the $\log_2$ ratio of expressed genes 24 hours after methanol induction compared to the expression levels before induction.

are shown Fold change is calculated as the $\log_2$ ratio of expressed genes 24 hours after methanol induction compared to the expression levels before induction.

4.3.4 Heterologous expression and host protein biogenesis demands

We next compared host protein synthesis in GS115 Mut^{+} and GS115 Mut^{S} ALB cultures. Prior to heterologous induction, protein synthesis rates per gene are highly conserved between GS115 Mut^{+} and GS115 Mut^{S} ALB as they have a Pearson's correlation of 0.97 (**Fig. 4-5a, Supplementary Fig. 4-3**). However, expression diverges significantly over time between the two strains as they show Pearson's correlations of 0.94 and 0.32 after 3 and 24 hours of methanol induction.

Host cell proteins can translocate into the ER using either co-translational or post-translational pathways, while heterologous proteins typically use co-translational pathways. In *K. phaffii*, we estimate 56 protein products to enter the ER post-translationally and 931 protein products to enter the ER co-translationally. Before induction, about 12% of nascent chains in each strain were predicted to enter the secretory pathway, with a similar distribution between co- and post-translational entry into the ER. This distribution aligns with findings from our previous study on *K. phaffii*[24]. After 24 hours of methanol induction, this increased to around 17% for both strains. However, the ratio of nascent chains entering the ER co- and post-translationally greatly diverges between strains (**Supplementary Fig. 4-4**). Post-translational translocation decreases significantly in GS115 Mut^{+} 24 hours after induction, with nearly an 85% reduction, while it remains unchanged in GS115 Mut^{S} ALB. The total production of post-translationally translocated nascent chains by GS115 Mut^{S} ALB was 6.6 times higher compared to the control strain, with the majority of these nascent chains being extracellular (27 genes) and membrane proteins (16 genes) (**Fig. 5b**). Co-translational translocation, however, increases in both strains, with a 2.3-fold increase in GS115 Mut$^+$ and a 2-fold increase in GS115 Mut^S ALB. While co-translationally translocated proteins localized to the cell membrane experience the most increase for the HSA-producing strain, peroxisomal proteins in the control strain show the greatest increase, likely due to faster methanol consumption.

Additionally, Both co- and post-translationally translocated genes experience different expression levels between the two strains (**Fig. 4-6** and **Supplementary Fig. 4-5**). We investigated which host cell proteins might limit heterologous protein entry into the ER by sequestering Sec-translocons during different stages of expression in GS115 Mut^S ALB. To focus on proteins that could restrict bioproduction, we excluded ER-resident proteins from our "hit list" as their deletion could impair folding and secretion of heterologous proteins. For nascent chains that enter the ER co-translationally, a mixture of both membrane and secreted proteins represent those that are the most highly expressed. After methanol induction, the most highly differentially expressed proteins between the two strains that are co-translationally translocated are *ATO2*, *YDR134C*, *ERG3, GAL2*, *OLE1* and *BGL2. ATO2* is a putative transmembrane protein involved in export of ammonia and it is the paralog of *ADY2* in *S. cerevisiae*. *YDR134C* is a non-essential secreted protein involved in cell wall maintenance that is homologous to *S. cerevisiae*'s paralog of *CCW12.* ER-resident *ERG3* and *OLE1* are desaturases required for lipid biosynthesis and metabolism. *GAL2* is a non-essential membrane protein involved in carbohydrate import and acetate transport. BGL2 is endo-beta-1,3-glucanase and is a major protein of the cell wall involved in cell wall maintenance.

For nascent chains that enter the ER post-translationally, the most highly expressed proteins are secreted and remain conserved before and after induction. The most highly differentially, post-translationally translocated proteins are *SPI1*, *XP_002494332.1*, *SCV12161.1,* and *AOA65896.1*. These proteins are relatively small cell wall constituents, and they are predicted to localize extracellularly. By focusing on these highly differentially expressed genes under heterologous conditions we may be able to rationally engineer secretion as a complex phenotype.

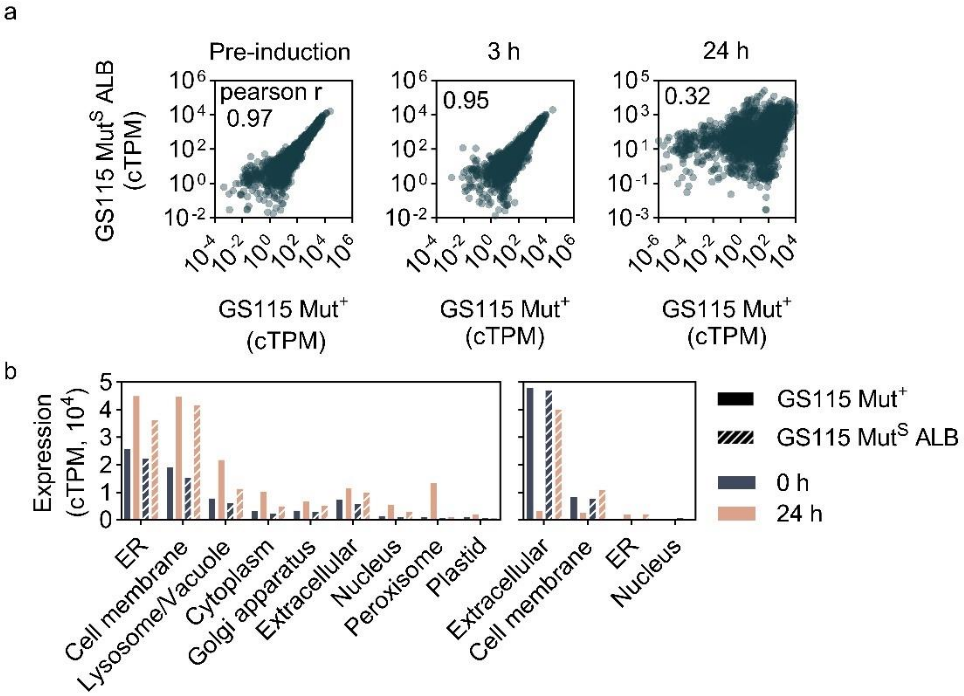


Fig. 4- 5. a) Divergence of translational landscape after heterologous expression. Prior to heterologous induction, nascent proteins produced per gene for GS115 Mut$^+$ and GS115 Mut^S ALB are highly correlated. After heterologous induction with methanol media, expression diverges between the two strains. **b)** Total nascent chain production of proteins translocating into the ER co- (left) and post-translationally (right) belonging to various subcategories. Co- and post-translationally translocated proteins with different localization predictions.

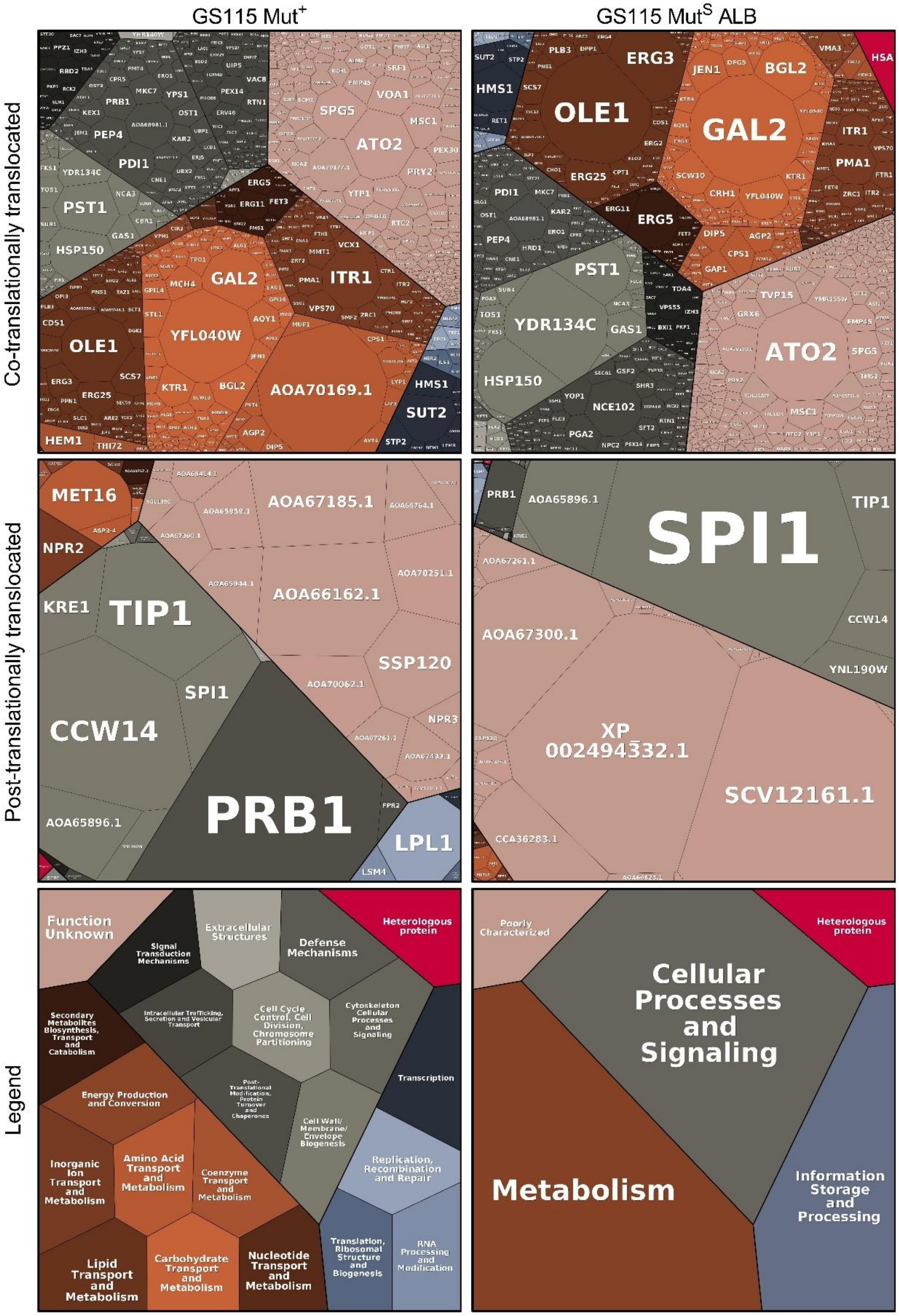

GS115 Mut$^+$
GS115 Mut^S ALB
Co-translationally translocated
Post-translationally translocated
Legend
PPZ1
RBD2
PRB1
MKC7
YPS1
PEP4
KEX1
OST1
PEX14
VAC8
RTN1
KAR2
PDI1
YDR134C
PST1
GAS1
HSP150
ERG5
ERG11
FET3
OLE1
GAL2
ITR1
MCH4
YFL040W
AOY1
VPS70
PMA1
VCX1
CDS1
SCS7
ERG3
ERG25
KTR1
BGL2
AGP2
DIP5
AOA70169.1
CPS1
HMS1
SUT2
STP2
HEM1
THI72
SPG5
VOA1
SRF1
MSC1
ATO2
PEX30
PRY2
YTP1
RTC2
PLB3
DPP1
JEN1
VMA3
HSA
CRH1
SCW10
GAP1
TVP15
GRX6
NCE102
YOP1
PGA2
HRD1
MET16
NPR2
KRE1
TIP1
SPI1
CCW14
AOA65896.1
PRB1
AOA67185.1
AOA66162.1
SSP120
NPR3
LPL1
LSM4
FPR2
AOA67300.1
YNL190W
XP 002494332.1
SCV12161.1
CCA36283.1
Function Unknown
Signal Transduction Mechanisms
Extracellular Structures
Defense Mechanisms
Heterologous protein
Secondary Metabolites Biosynthesis, Transport and Catabolism
Intracellular Trafficking, Secretion and Vesicular Transport
Cell Cycle Control, Cell Division, Chromosome Partitioning
Cytoskeleton Cellular Processes and Signaling
Transcription
Energy Production and Conversion
Post-Translational Modification, Protein Turnover and Chaperones
Cell Wall/ Membrane/ Envelope Biogenesis
Replication, Recombination and Repair
Inorganic Ion Transport and Metabolism
Amino Acid Transport and Metabolism
Coenzyme Transport and Metabolism
Lipid Transport and Metabolism
Carbohydrate Transport and Metabolism
Nucleotide Transport and Metabolism
Translation, Ribosomal Structure and Biogenesis
RNA Processing and Modification
Poorly Characterized
Heterologous protein
Cellular Processes and Signaling
Metabolism
Information Storage and Processing

Fig. 4- 6. Co and post-translational flux through the ER for GS115 Mut$^+$ and GS115 Mut^S ALB strains 24 hours after induction. Non-mitochondrial proteins are predicted to enter the secretory pathway co-translationally if they have greater than $\log_2$ membrane enrichment in YPD studies. Gene products are grouped by ontological function using COG scores predicted by EggNOG v5.0. Cell sizes are calculated using cTPM scores and represent relative quantities of nascent chains produced per gene. Tessellation plots are made using www.bionic-vis.biologie.uni-greifswald.de[48–50].

4.3.5 Ribo-seq-enabled rational engineering of *K. phaffii* to improve protein secretion

Knocking out highly expressed, non-essential genes sequestering the most resources related to protein secretion under heterologous conditions is a rational approach for genetic modification. We next investigated the influence of single and multiple gene knockouts on HSA secretion using the CRISPR-Cas9 system. Based on the Ribo-seq analysis, we initially selected *GAL2*, *YDR134C*, *BGL2*, *SPI1*, *XP_002494332.1*, *SCV12161.1*, and *AOA65896.1* for validation experiments. However, *OLE1* and *ERG3,* being ER-resident proteins and required for metabolism, were removed from our validation experiments. Additionally, we were not able to knock out *SPI1* and *XP_002494332.1* even with multiple sgRNAs. SPI1 has been identified as essential for growth based on our recent genome-wide CRISPR-Cas9 growth screen [51]. As a result, the final list for further validation included the remaining five genes.

To measure HSA secretion, single or combined gene knockouts were introduced into GS115 Mut^S ALB cells. The cells were first cultivated in glycerol media for biomass accumulation, then transferred to BMMY media with 0.5% methanol for five days, with daily addition of 0.5% methanol. After this period, the supernatant was collected, concentrated for better visualization of bands on SGS-PAGE gels, and the HSA band volume was measured. (**Fig. 4-7a**). For more accurate secretion measurements, first the band volume of HAS on each lane was normalized to the OD_{600} of the sample in order to be able to differentiate between the impact of gene knockouts on the fitness of the cell and HAS secretion levels. HSA band volume was also normalized to the band volume of blue fluorescent protein (BFP) added equally to each well to standardize the band volumes enabling a more accurate comparison between lanes. The relative normalized band volume was calculated as the ratio of normalized band volume of each sample to the same of the control strain (GS115 Mut^S ALB) (**Fig. 4-7b and c**).

We first investigated the influence of single knockouts on HSA secretion. It was observed that a single knockout in *GAL2*, *BGL2*, and *YDR134C* genes was able to increase HSA secretion 28, 18, and 26%, respectively, compared to the control strain. Whereas single knockouts in *SCV12161.1* and *AOA65896.1* genes did not statistically change HSA secretion.

We next sought out the influence of double gene knockouts on HSA secretion. The addition of *BGL2* knockout to strains with either *YDR134C* or *GAL2* single knockout was associated with statistically significant lower levels of HSA secretion (0.82- and 0.92-fold secretion in GS115 Mut^S ALB *ydr134c bgl2* and GS115 Mut^S ALB *gal2 bgl2* double-knockouts, respectively, compared to the control strain). Furthermore, HSA secretion did not change with the addition of *AOA65896.1* knockout to the GS115 Mut^S ALB *gal2* strain, and it resulted in lower secretion levels in the GS115 Mut^S ALB *ydr134c* strain. The highest increase in HSA secretion was observed with a triple-knockout in *GAL2, AOA65896.1* and *YDR134C* gene which led to a 35% improvement of

HSA secretion compared to the control strain. Other studied combinations or additional knockouts were not able to improve secretion further. In conclusion, our study demonstrates the use of ribosome profiling and next generation sequencing to study the nascent chain protein production under heterologous conditions. These data can be used to rationally engineer complex phenotypes such as protein secretion.

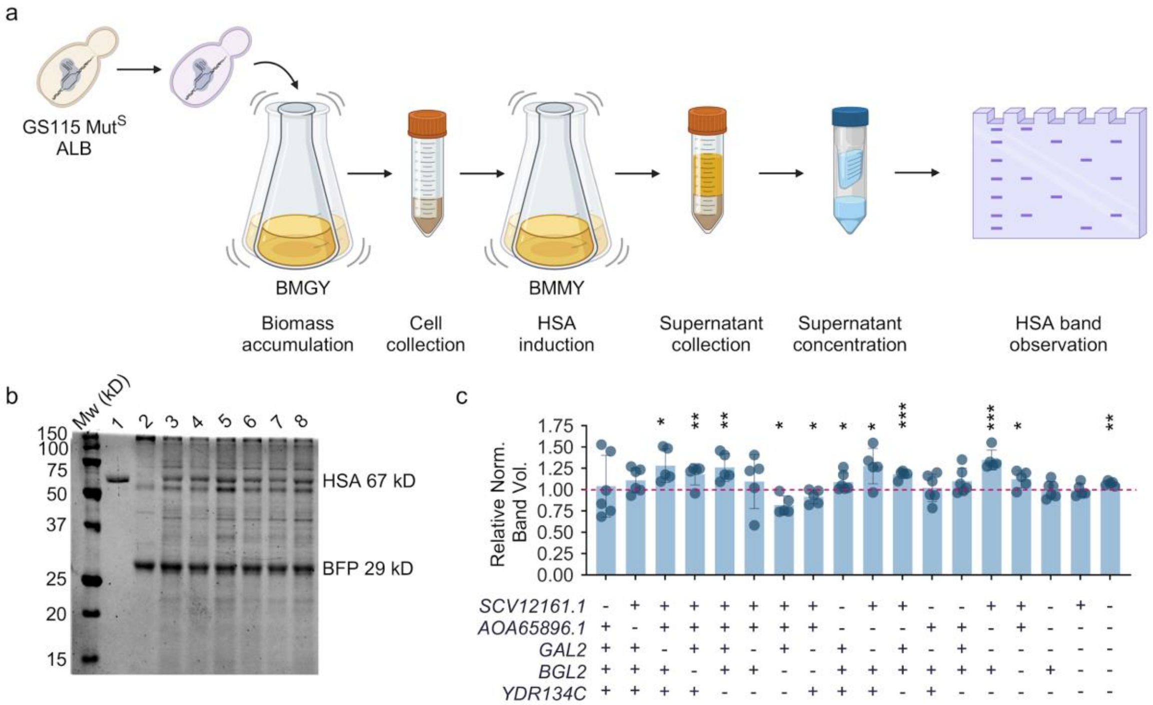


Fig. 4- 7. a) Experimental workflow of rational engineering of *K. phaffii* cells enabled by Ribo-seq. Five genes with high expression levels sequestering the early secretory pathway were identified by Ribo-seq. Subsequently single and multiple knockout strains were produced using the CRISPR-Cas9 system in the GS115 Mut^S ALB strain. Strains were grown in BGMY media overnight and then transferred to BMMY media for HSA induction for five days. The supernatant was extracted, concentrated using filters with a 50 kDa cutoff, and mixed with Blue Fluorescent Protein (BFP) as the standard allowing for a more accurate HSA qualification between lanes. HSA band volume was normalized and measured on SDS-PAGE gels to compare heterologous protein secretion. **b)** Secreted proteins into the media from GS115 Mut^S ALB cells with single knockouts visualized on SDS-PAGE gels. Lanes from left to right: protein standard, pure recombinant HSA (1), GS115 Mut$^+$ strain (2), GS115 Mut^S ALB (3), GS115 Mut^S ALB *gal2* (4), GS115 Mut^S ALB *aoa65896.1* (5), GS115 Mut^S ALB *bgl2* (6), GS115 Mut^S ALB *scv12161.1* (7), GS115 Mut^S ALB *ydr134c* (8). **c)** Relative normalized HSA band volume in strains with either single or multiple gene knockouts. Bars represent the ratio of secreted normalized HSA from each knocked out strain compared to the base GS115 Mut^S ALB strain. HSA band volume was normalized to the OD_{600} of each sample and the band volume of blue fluorescent protein (BFP) added equally to each lane. Experiments were done in six biological replicates and the bars show the mean of the replicates. * $p < 0.05$, ** $p < 0.005$, and *** $p < 0.005$; one-tailed paired t-test).

4.4 Discussion

We hypothesize that cells' biogenetic machinery has co-evolved under the demands of their own proteome. Gaining a deeper understanding of how a production chassis operates under heterologous conditions could offer unique insights for improving bioproduction [52]. In *K. phaffii*, transcriptomic studies have identified gene targets that are differentially expressed during heterologous production and could be overexpressed to enhance bioproduction [53]; many of these genes like vacuolar *VMA3*, golgi *COG6*, and COPII vesicle *SEC31* were also differentially expressed between GS115 Mut^+ and GS115 Mut^S ALB in our study. Other studies in *K. phaffii* have shown that increased heterologous expression in lower temperature conditions were due to lesser expression of genes involved in the UPR, rather than increased heterologous expression [54]. We aimed to use -omics based approaches so that we may identify bottlenecks that may hinder bioproduction in our conditions [55–57]. Notably, recent research has focused on understanding the biosynthetic demands of the host proteome in Chinese hamster ovary cells under heterologous conditions. This approach involves selectively depleting non-essential mRNAs, which has been shown to enhance growth rates, improve product quality, and increase protein secretion [58,59].

In *K. phaffii*, industrial bioproduction typically relies on glycerol-based media for cell growth and methanol-based media for heterologous induction [5–7]. We compared host protein synthesis between GS115 Mut$^+$ and GS115 Mut^S ALB under these conditions. The most significant difference between the two strains is that GS115 Mut$^+$ and GS115 Mut^S ALB metabolize methanol at different rates. Methanol is utilized as a substrate for energy production using alcohol oxidase (AOX) generated by *AOX1* and/or *AOX2* with *AOX1* having the majority of the AOX activity[41,42,60]. In Mut^S strains, AOX production relies solely on *AOX2* expression as *AOX1* is disrupted. Many heterologous proteins are expressed and glycosylated using the *AOX1* promoter [61]. For these proteins, higher production titers are observed in the slow growing Mut^S strain than the fast growing Mut$^+$ strain [44,61]. Observing protein synthesis of both strains under these conditions provides insight into possible strategies for strain engineering.

During the first step in the methanol utilization pathway, peroxisomal AOX generates high levels of H_2O_2 during methanol catalysis. We find that methanol metabolism leads to increased expression of *YAP1* and *GSH2*, where Yap1 is a required transcription factor for *GSH2* which expresses glutathione in the glutathione redox system [62,63]. These findings are accompanied by increased expression of genes like *GCN4* and *GLN1* whose products import amino acids constituent of thiol-containing peptides involved in redox reactions [64,65]. While RNA-seq has been used to study oxidative stress responses proceeding methanol metabolism [66], Ribo-seq is a more sensitive and appropriate tool for quantifying protein levels as oxidative stress increases the frequency of post-transcriptional modifications [66–68]. For instance, RNA-seq finds *DAS1* and *DAS2* equally expressed after methanol induction [44] while Ribo-seq shows *DAS2* to be more highly expressed than *DAS1*. Ribo-seq also reveals translational dynamics that indicate methanol induced oxidative stress responses. At many loci, we observe translation initiation events upstream ORFs at 5'UTRs after methanol induction similar to other studies of H_2O_2 treated yeast cultures [68]. As well, our analyses are congruent with previously observed reductions in protein synthesis rates consequential to oxidative stress [69] as we find decreased expression of genes encoding ribosomal proteins (e.g., *RPS23B* and *RPL3*) and increased expression of genes encoding RNA-binding proteins thought to stabilize slowly translating transcripts from degradation (e.g., *NRD1*, *NAB3*, and *PAB1*) [70]. Together, we find GS115 Mut^S ALB less affected by methanol induced oxidative stresses than GS115 Mut^+, likely due to lesser AOX expression and subsequently lesser H_2O_2 generation. Additionally, we found that GS115 Mut^S ALB shows lower overall expression levels for genes involved in the UPR and ERAD compared to GS115 Mut^+, similar to another study[45]. It has been reported that cells cultured in glycerol media experience higher levels of UPR activation. This suggests that the elevated UPR activity might help alleviate the burden associated with the strong induction of the recombinant product[71].

We sought to understand how heterologous production affects early secretory trafficking of host cell proteins. Highly expressed host cell proteins that enter the early secretory pathway sequester biogenesis machinery that are limited in number and processivity which may limit heterologous secretion. Host cell proteins may enter the early secretory pathway co-translationally or post-translationally depending on their protein sequence features and translational dynamics. The majority of proteins using co-translational pathways are SRP-dependent and contain hydrophobic targeting sequences like transmembrane domains [72], N-terminal signal sequences [73], and/or glycosylphosphatidylinositol (GPI) anchors [74]. SRP is often pre-recruited to the ribosome nascent chain complex (RNC) [75,76] and thus binds quickly to an emerging hydrophobic targeting sequence [24,77]. Some proteins containing N-terminal signal sequences complete translation before they have time to reach the ER [78] and are instead translocated post-translationally [79]. These proteins typically contain few amino acids [24]. For proteins that do not contain an N-terminal signal sequence, GPI anchors at the carboxyl terminus allow them to translocate post-translationally in an SRP-independent manner [74]. As proteins with similar features can enter the ER co- and post-translationally, we used protein sequence features as well as Ribo-seq reads from cytosolic and membrane bound ribosomes in GS115 Mut$^+$ cultured in YPD [24] to predict their trafficking pathways. The assumption that proteins translocate similarly under heterologous conditions relies on two previous observations: *K. phaffii*'s secretome does not change with different carbon substrates [80] and that proteins' ER translocation routes are contingent on their sequence features and constituent number of amino acids [24,77].

In comparing GS115 Mut$^+$ and GS115 Mut^S ALB, the percentage of nascent chains predicted to enter the ER similarly increased after 24 hours of methanol induction. However, a significantly greater number of cell wall and membrane nascent chains entered the ER for GS115 Mut^S ALB. The molecular organization of the cell wall is dynamic. The mechanical strength of the cell wall is largely due to the inner layer consisting of β 1,3-glucan and chitin [81]. The outer layer of the cell wall consists of glycosylated mannoproteins covalently linked to the β 1,3-glucanchitin network directly or disulfide bound to other cell wall proteins. Cell wall mannoproteins affect stability and resistance to stress [82–84]. As the extracellular and membrane proteins that largely constitute differences between strains are not those shown to be inductively expressed from oxidation [85], it would appear that reorganization of GS115 Mut^{S} ALB's cell wall is instead consequent to stresses imposed by heterologous secretion. Therefore, the most highly expressed cell wall and membrane proteins entering the ER at different stages of heterologous expression offer novel insights for improving secretion.

While the diversity and number of post-translationally translocated nascent chains do not appreciably change after induction, their expression levels are amongst the highest observed. Of this group, *SPI1* is consistently one of the most highly expressed proteins in *K. phaffii* [24,86,87]. For co-translationally translocated gene products, *GAL2*, *YDR134C*, and *BGL2* are amongst the most highly differentially expressed between strains. As galactose is preferentially incorporated into cell wall glucan over glucose [88], we speculate that *GAL2* is upregulated to increase the overall expression of cell wall mannoproteins. Since the carbon source used in these experiments does not contain galactose, *GAL2* is a good candidate for strain engineering under heterologous conditions. *YDR134C* and BGL2 are both major cell wall proteins. *YDR134C* is a heavily glycosylated, nonenzymatic GPI-anchored protein involved in the cell wall organization and is the paralog of *CCW12* in *S. cerevisiae*. *BGL2* is involved in cell wall maintenance and incorporating newly synthesized mannoprotein molecules into the cell wall. The effect of cell wall proteins on recombinant protein secretion is not well studied. A study found the disruption of the cell wall mannoprotein *CWP2* increased cell wall permeability and cellobiohydrolase heterologous secretion [89–91]. Another study found the inactivation of *DFG5*, *YPK1*, *FYV5*, *CCW12* and *KRE1* increased cellulolytic enzyme β-glucosidase secretion and surface display in *S. cerevisiae* [92].

4.5 Materials and Methods

4.5.1 Strains and culture conditions

All strains used in this work are presented in **Supplementary Table 4-2**. Assays were performed using GS115 Mut$^+$ and GS115 Mut^S ALB (Pichia expression kit, Life Technologies, 2014). For each biological replicate, 200 mL liquid cultures of BMGY (1 % yeast extract, 2 % peptone, 100 mm potassium phosphate pH 6.0, 1.34 % YNB, and 1 % glycerol) were grown to an OD_{600} nm of 5 at 30 °C with shaking in baffled 2 L flasks. Of this culture, 100 mL were harvested by vacuum filtration through a 0.8 µm filter. Immediately after filtering, cells were scraped off the filter using a chilled scoopula and submerged in a 50 mL conical tube containing liquid nitrogen. The remaining liquid cultures were split into two 50 ml conical tubes and were pelleted via centrifugation. Supernatant was removed from each 50 ml conical tube. The cell pellet of one 50 ml conical tube was gently resuspended with 40 mL BMMY without methanol (1 % yeast extract, 2 % peptone, 1.34 % YNB, and 100 mm potassium phosphate pH 6.0). Resuspended culture was used to resuspend the cell pellet in the second 50 ml conical tube. Resuspended cultures were equally divided into two 280 mL cultures of BMMY without methanol in 2 L baffled flasks for a final volume of 300 mL for each sample.

Methanol was added at 0.5 % to each of the baffled flasks for *AOX1* induction. Flasks were allowed to shake at 30 °C and were collected in the manner described above three and twenty-four hours after methanol induction. Lysis buffers (50 mM MOPS, 25 mM KOH, 100 mM KOAc, 2 mM MgOAc, 1 mM DTT, and 1 % Triton X-100) for each sample were frozen by adding 2 mL dropwise to a 50 mL conical tube containing liquid nitrogen. For each sample, frozen cells were mixed with 2 mL frozen lysis buffer. Cell fractions were pulverized for 2 min in a 50 mL ball mill chamber with a single 2 cm steel ball (Retsch) and collected in 50 mL conical tubes. After thawing, lysates were centrifuged at 18,000 g for 10 min. Supernatants were transferred to 1.5 mL conical tube and were further clarified by centrifugation at 23,000 g for 20 min.

For yeast transformation experiments, cells were grown in 100 ml YPD at 30 °C and 225 RPM and were transformed with plasmids expressing sgRNAs. Transformants were grown in 2 ml of YPD supplemented with 800 μg/ml of G418 in 14 ml polypropylene tubes. For HSA secretion quantification, cells with the knockout of interest were grown in 200 ml of BMGY supplemented with 1% glycerol in 1 L shake flasks for biomass accumulation until OD_{600} = ~ 6. Next, cells were transferred to 20 ml of BMMY with 0.5% methanol for HSA induction in 100 ml shake flasks. Cells were grown in BMMY media for five days with daily supplementation of 0.5% methanol (v:v).

4.5.2 Ribo-seq

Lysed samples were nuclease digested using 40 U of Ambion RNase A for one hour at room temperature. Digested samples were layered on a 10 % to 50 % sucrose gradient prepared in 50 mM Tris pH 7.5, 200 mM NaCl, and 2 mM MgOAc case using a Gradient

Master (Biocomp). Gradients were centrifuged at 39,000 RPM for 2.5 h in a TH-641 rotor (Thermo). After centrifugation, gradients were fractionated using a Piston Gradient Fractionator (Biocomp) and monosome peaks were retained. Total RNA was extracted using a standard phenol-chloroform method and alcohol precipitated. Ribosome protected footprints 18 nt to 34 nt were resolved and excised using 15 % polyacrylamide TBE-urea gel. RNA was collected from excised gel fragments using RNA gel extraction buffer (300 mM NaOAc, 1 mM EDTA, and 0.25 % SDS), precipitated, and resuspended in water containing 20 U mL^{-1} SUPERase·In.

Purified fragments were then dephosphorylated by incubating 2 μL 1 M RNA sample with 2 μL RNase free water, 0.5 μL SUPERase·In RNase Inhibitor, 0.5 μL T4 Polynucleotide Reaction Buffer (PNK), and 0.5 μL T4 Polynucleotide Kinase at 37 °C for 1 h. Dephosphorylated samples were linker ligated with adapter sequences by incubating with 3.5 μL 50 % PEG-8000, 0.5 μL 10X T4 RNA Ligase Reaction Buffer, 0.5 μL 10 μM adenylated linkers and 0.5 μL T4 Rnl2(tr)k277Q at 30 °C for 4 h. Linker-ligated samples were concentrated via isopropanol precipitation and resolved using 15 % TBE-urea polyacrylamide gel. Imaged samples were diluted and pooled to equivalent concentrations by their relative pixel intensities calculated from BioRad imaging software after overnight extraction in RNA gel extraction buffer.

Ligated and purified samples were rRNA depleted using streptavidin-coated magnetic beads from the Ribo-Zero rRNA Removal Kit as recommended by the manufacturer. Depleted samples were precipitated, resolved using 15 % TBE-urea polyacrylamide gel, and extracted as previously described.

RNA was reverse transcribed by adding 2 µL reverse transcription primer to 10 µL sample and incubating at 65 °C for 5 min to denature. Denatured sample was then incubated with 4 µL 5X First Strand Buffer, 1 µL 10 mM dNTPs, 1 µL 10 mM DTT, 1 µL 20 U µL −1 SUPERase·In and 1 µL 200 U μL^{-1} SuperScript II Reverse Transcriptase at 50 °C for 30 min using thermal block. After incubation, sample was hydrolyzed by adding 2.2 µL 1 M NaOH and then incubated at 70 °C for 20 min using thermal block. 28 µL RNAse free water was added to reverse transcription mixture (~50 µL total) and concentrated using Oligoclean and Concentrator Kit. Concentrated RNA was then purified of reverse transcription primers using 12 % TBE-urea polyacrylamide gel. RNA from gel slices was extracted using the method previously described. Extracted precipitants were resuspended in 11 µL 1:1000 SUPERase·In.

Single stranded cDNA samples were circularized by incubating 11 µL sample in 2 µL CircLigase II 10x Reaction Buffer, 1 µL 50 mM $MnCl_2$, 1 µL ATP, 4 µL 5 M Betaine, and 1 µL 100 U μL^{-1} CircLigase II ssDNA Ligase at 60 °C for 3 h on thermal block. The circularization process was inactivated by incubating the sample at 80 °C for 10 min on a thermal block.

Circularized samples were rRNA depleted, again, using probe-directed degradation via double stranded nuclease (DSN) [50,102] . Depletion probes were designed using rRNA aligned Ribo-seq reads collected from GS115 Mut^S ALB cultured in BMGY before methanol induction. Circularized samples (10 µL) were incubated with 4 µL 4x hybridization buffer, 1 µL 4x depletion probes at 200 µM, and 1 µL water. Mixture was denatured at 98 °C for 2 min and allowed to slowly anneal at 65 °C for 5 h. Double stranded rRNA fragments were enzymatically degraded by adding 2 μL 10x DSN master buffer, 1 μL DSN storage buffer, and 1 μL DSN before incubation at 65 °C for 25 min. Reaction was stopped by adding 20 μL 10 mM EDTA to DSN depleted sample mix. Samples were then purified using AMPure XP beads. After DSN treatment, samples were digested using Exonuclease I to degrade linearized DSN degraded DNA fragments as these may contain regions complementary to PCR amplification primers. Samples were again purified using AMPure XP beads.

Circularized samples were PCR amplified for 16 cycles using a 50 μL reaction mixture (10 μL Q5 Reaction Buffer, 1 μL 10 mM dNTPs, 2.5 μL 10 μM forward primer, 4 μL circularized DNA sample, 0.5 μL Q5 High Fidelity DNA Polymerase and 29.5 μL RNAse free water) divided into 5 x 10 μL aliquots. Amplified sample was resolved using 10 % non-denaturing TBE polyacrylamide gel and extracted using previously described method. Libraries were quantified using Qubit 2.0 Fluorometer and sequenced using Illumina NextSeq.

4.5.3 Mapping of ribosome-protected reads to codons

Sequenced reads were trimmed and demultiplexed in an error-tolerant way using Cutadapt 2.3 [103]. Reads were computationally rRNA subtracted by aligning them to *Komagataella pastoris* GS115 genomic rRNA using HISAT2 [104,105]. Subtracted reads were mapped to the genome for *Komagataella pastoris* GS115 [106] using HISAT2. Sequence alignment map (SAM) files were converted to sorted and indexed binary alignment map (BAM) files using Samtools and only included reads of high mapping quality [107,108]. Genomic alignments were loaded into R using the GenomicAlignments package from Bioconductor [109]. Genomic alignment ranges were converted to their 3' end positions before determining p-site offsets. P-site offsets were determined using the existing genome annotations [13] and the RiboProfiling package in Bioconductor [110]. Genomic alignment objects were used with p-site offsets to generate reads per codon per gene (RPCPG) data tables.

4.5.4 Masking reads of ambiguously mapped codons

Codon masks were created by first parsing the coding sequence annotation file associated with the reference genome into a FASTA file simulating every possible 28 NT combination (approximate length of a ribosome protected mRNA fragment). This FASTA file was then aligned to the reference genome twice, once to only include reads with mapping quality greater than or equal to 60 (unambiguously assigned), and another to include all reads (ambiguously assigned). Both alignment files were used to generate RPCPG data tables using methods previously discussed. The unambiguously assigned reads were subtracted from ambiguously assigned reads and codons with a nonzero difference were included in the mask. The first and last five codons in genes' open reading frames were masked to correct for variable read quality at the beginning and ending of transcripts inherent to Ribo-seq [111].

4.5.5 Normalization and differential expression analysis

Read counts were normalized at the codon level using a metagene correction strategy previously discussed [13] with some modification. Reads for the first 500 codon positions at the 5'end of all transcripts was scaled by their respective codon-specific normalized metagene values. Normalized metagene values were calculated for all codons in all ORFs and applied in the following manner: positions 1 to 100 were normalized with a rolling mean with a window of 10 codons and positions 100 to 500 were normalized with a rolling mean with a window of 100 codons. Scaled reads per gene were calculated as the sum of a gene's scaled codon reads (codon positions less than or equal to 500) and unscaled codon reads (codon position greater than 500).

Gene read count thresholds were calculated using an adapted method [18]. First, we summed the scaled reads per gene for each gene between biological replicates. Each gene was grouped into 1 of 50 quantiles using the probabilistic distribution of the summed scaled read counts between replicates. In calculating the read count threshold for one replicate, the replicate's scaled reads per gene were normalized by the summed read count for their respective bin. The standard deviation of normalized fractions across each bin were plotted against the summed read value for each bin. Read count thresholds were calculated as the knee-point in the exponential regression for this plotted curve. This process was repeated to calculate unique read count thresholds for each biological replicate. Read count thresholds were linearly regressed on the total reads for that replicate to conservatively predict thresholds for all samples.

Scaled and filtered reads were normalized by their pseudo gene lengths (theoretical gene length minus number of masked codons) and sequencing depth to give corrected transcripts per million (cTPM). Genes were described as significantly expressed if their cTPM values were among the upper 75th percentile of cTPM values for that sample. For differential expression, genes were described by their fold enrichment between samples if both samples had scaled read counts above their respective read count thresholds. Fold enrichment scores were also used to quantify differential expressions between groups of genes categorized by their ontological function. In cases where only one sample showed read counts above their respective read count threshold, genes were simply described as enriched.

4.5.6 Classification of ORFs

Open reading frames for each gene were characterized using various prediction softwares: clusters of orthologous groups were predicted using EggNOG 4.5 [112], subcellular localization was predicted using DeepLoc [113], signal sequences were predicted using SignalP 5.0 [114], transmembrane domains were predicted using TOPCONS [115], and GPI-anchors were predicted using predGPI [116]. ER-targeting classifications were made for each gene using Ribo-seq data sets from subcellularly-fractionated GS115 Mut$^+$ collected during log phase growth in YPD [13]. These data sets revealed expressions from translating ribosomes in the cytosol and on the membrane of the ER and mitochondria. The $\log_2$ ratio of cTPM scores for genes in membrane and cytosolic fractions were used to generate membrane enrichment scores. Membrane enrichment scores were used with protein sequence predictions to determine which gene products are translocated into the ER co- and post-translationally. Co-translationally translocated genes had greater than 2-fold membrane enrichment. This classification was broader to include membrane proteins (containing more than two extracytoplasmic transmembrane domains), secreted proteins (containing an N-terminal signal sequence and at most one transmembrane domains near the C-terminus), and proteins without these features that may target the ER using mechanisms involving the 3'UTR. Post-translationally translocated genes show less than 2-fold membrane enrichment and contain a predicted N-terminal signal sequence and less than or equal to one transmembrane domain or a GPI-anchor at the C-terminus. Genes products that met these criteria were filtered to remove those that were predicted to localize to mitochondria.

4.5.7 Rational strain engineering of *K. phaffii* for improved HSA secretion

To validate candidate genes identified by Ribo-seq to rationally improve HSA secretion, *GAL2*, *BGL2*, *YDR134C, SCV12161.1,* and *AOA65896.1* were chosen and knocked out using CRISPR-Cas9 system (**Supplementary Table 4-3**). The D-227 vector containing *CAS9* and a sgRNA expression cassette, kindly donated from Dr. Christopher Love Lab at Massachusetts Institute of Technology [117], was used for gene knockout. The plasmid was digested with New England BioLabs (NEB) BbVCI restricition enzyme according to the manufacturer's instructions, and sgRNAs were cloned following our lab's established protocol [118]. Primers for sgRNA cloning were obtained from Integrated DNA Technology (IDT). All cloning was performed with NEBuilder® HiFi DNA Assembly Master Mix (NEB) according to the manufacturer's instructions. Successful cloning of the sgRNA fragment was confirmed by Sanger sequencing.

After verifying the sgRNA cloning, plasmids were transformed into GS115 Mut^S ALB strain using a modified electroporation method [24]. Transformants were grown in 4 ml of selective media (YPD + 800 µg/ml G418) for two days before being transferred to 2 ml fresh selective media, where they were allowed to grow for an additional two days before plating them on selective plates (YPD, 800 µg/ml G418, and 2% agar). 8 single colonies were randomly picked, the targeting gene was PCR-amplified, and Sanger sequencing was used to verify the successful knockout of the gene (**Supplementary Table 4-4**).

For HSA quantification, the strain of interest was grown in 200 ml of BMGY until the OD_{600} reached ~ 6. Next, 2×10^{10} cells were transferred to BMMY for HSA induction. Cells were grown in the BMMY media for five days at 30°C and 225 RPM with daily supplementation of 0.5% methanol. After this period, cells were centrifuged at 3000×g for 5 minutes. 100 µl of the supernatant was mixed with 400 µl of Nuclease Free water and was loaded on Amicon® Ultra Centrifugal Filter with a 50 kDa cutoff. Samples were centrifuged at 14000×g for 8 minutes and were washed once again with 500 µl of nuclease free water at 14000×g for 8 minutes. The concentrated supernatant was then recovered and 1 µl of concentrated sample was mixed with 2 µl blue fluorescent protein (BFP), 5 µl of nuclease free water, and 2 ul of 5X SDS Loading buffer. The mixture was heated at 100 °C for 5 minutes to denature the proteins and was loaded on SDS-PAGE gels. To quantify the volume of the HSA band, Bio-Rad's Molecular Imager® ChemiDoc™ XRS System was used. The device's software was used to identify HSA and BFP bands and to measure their size.

Blue fluorescent protein (BFP) was expressed from plasmid pCRG068 (Supplementary Data) transformed into TOP10 *E. coli* cells. Ni-NTA affinity chromatography was used to purify the polyhistidine-tagged BFP. Briefly, *E. coli* transformants were grown in 100 ml of LB media supplemented with 100 mg/L of ampicillin for 24 hours at 37 °C and 225 RPM. Samples were centrifuged, the supernatant was discarded, and cells were distributed in 10 ml of the lysis buffer (50 mM Tris (pH=7.5), 500 mM NaCl). Cells were then sonicated, and the remaining cell pellets were separated from the supernatant using centrifugation. Thermo Scientific disposable plastic columns were packed with 2 ml of Ni-NTA resins according to the manufacturer's instructions. And the supernatant was loaded on the Ni-NTA resins. After all the supernatant was loaded, the resins were washed three times with the wash buffer (50 mM NaH2PO4 (pH 8.0) and 0.5 M NaCl) and the BFP protein was eluted by gradually adding 2 ml of the elution buffer (3 M Imidazole, 500 mM NaCl, 20 mM Sodium Phosphate Buffer pH = 6.0). To concentrate the purified BFP, Amicon® Ultra Centrifugal Filter with a 10 kDa cutoff was used according to the protocol explained before.

4.6 Conclusion

Heterologous protein production is a complex process that relies on the limited resources of the translational and secretory machinery. Through next-generation sequencing and ribosome profiling, we have gained valuable insights into the metabolic and secretory demands of the methylotrophic yeast *Komagataella phaffii* under heterologous conditions. Our findings revealed host protein flux changes through the endoplasmic reticulum in response to heterologous protein production. We also found that the variety and levels of host proteins entering the secretory pathway are unique to different stages of heterologous expression. Application of this tool allowed us to find highly expressed, non-essential gene targets that significantly consume resources in the early secretory pathway during methanol induction. Employing the CRISPR-Cas9 system, we performed single or combined gene knockouts to investigate the influence of these genes on rationally improving recombinant protein secretion in *K. phaffii*. Our investigation resulted in the identification of a triple knockout leading to a 35% improvement of human serum albumin (HSA) secretion from this industrially relevant yeast. These findings offer valuable insights for process optimization and strain engineering in industrial applications.

References

1. Sanchez-Garcia, L. *et al.* Recombinant pharmaceuticals from microbial cells: a 2015 update. *Microb. Cell Fact.* **15**, 33 (2016).
2. Love, K. R., Dalvie, N. C. & Love, J. C. The yeast stands alone: the future of protein biologic production. *Curr. Opin. Biotechnol.* **53**, 50–58 (2018).
3. Biologics Market Size, Share, Trends & Growth Report, 2033. https://www.novaoneadvisor.com/report/biologics-market.
4. Bernauer, L., Radkohl, A., Lehmayer, L. G. K. & Emmerstorfer-Augustin, A. Komagataella phaffii as Emerging Model Organism in Fundamental Research. *Front. Microbiol.* **11**, 607028 (2020).
5. Liang, S. *et al.* Comprehensive structural annotation of Pichia pastoris transcriptome and the response to various carbon sources using deep paired-end RNA sequencing. *BMC Genomics* **13**, 738 (2012).
6. Cereghino, J. L. & Cregg, J. M. Heterologous protein expression in the methylotrophic yeast Pichia pastoris. *FEMS Microbiol. Rev.* **24**, 45–66 (2000).
7. Gasser, B., Maurer, M., Gach, J., Kunert, R. & Mattanovich, D. Engineering of Pichia pastoris for improved production of antibody fragments. *Biotechnol. Bioeng.* **94**, 353–361 (2006).
8. Wang, G., Huang, M. & Nielsen, J. Exploring the potential of Saccharomyces cerevisiae for biopharmaceutical protein production. *Curr. Opin. Biotechnol.* **48**, 77–84 (2017).
9. Mattanovich, D., Gasser, B., Hohenblum, H. & Sauer, M. Stress in recombinant protein producing yeasts. *J. Biotechnol.* **113**, 121–135 (2004).
10. Salari, R. & Salari, R. Investigation of the Best Saccharomyces cerevisiae Growth Condition. *Electron Physician* **9**, 3592–3597 (2017).
11. Narendranath, N. V. & Power, R. Relationship between pH and medium dissolved solids in terms of growth and metabolism of lactobacilli and Saccharomyces cerevisiae during ethanol production. *Appl. Environ. Microbiol.* **71**, 2239–2243 (2005).
12. Karbalaei, M., Rezaee, S. A. & Farsiani, H. Pichia pastoris: A highly successful expression system for optimal synthesis of heterologous proteins. *J. Cell. Physiol.* **235**, 5867–5881 (2020).
13. Gustafsson, C., Govindarajan, S. & Minshull, J. Codon bias and heterologous protein expression. *Trends Biotechnol.* **22**, 346–353 (2004).
14. Mori, A. *et al.* Signal peptide optimization tool for the secretion of recombinant protein from Saccharomyces cerevisiae. *J. Biosci. Bioeng.* **120**, 518–525 (2015).
15. Bae, J.-H. *et al.* An Efficient Genome-Wide Fusion Partner Screening System for Secretion of Recombinant Proteins in Yeast. *Sci. Rep.* **5**, 12229 (2015).
16. Huang, M., Wang, G., Qin, J., Petranovic, D. & Nielsen, J. Engineering the protein secretory pathway of Saccharomyces cerevisiae enables improved protein production. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E11025–E11032 (2018).
17. Payne, T. *et al.* Modulation of chaperone gene expression in mutagenized Saccharomyces cerevisiae strains developed for recombinant human albumin production results in increased production of multiple heterologous proteins. *Appl. Environ. Microbiol.* **74**, 7759–7766 (2008).
18. de Ruijter, J. C., Koskela, E. V. & Frey, A. D. Enhancing antibody folding and secretion by tailoring the Saccharomyces cerevisiae endoplasmic reticulum. *Microb. Cell Fact.* **15**, 87 (2016).

19. Hansen, H. G., Pristovšek, N., Kildegaard, H. F. & Lee, G. M. Improving the secretory capacity of Chinese hamster ovary cells by ectopic expression of effector genes: Lessons learned and future directions. *Biotechnol. Adv.* **35**, 64–76 (2017).
20. Valkonen, M., Penttilä, M. & Saloheimo, M. Effects of inactivation and constitutive expression of the unfolded- protein response pathway on protein production in the yeast Saccharomyces cerevisiae. *Appl. Environ. Microbiol.* **69**, 2065–2072 (2003).
21. Zahrl, R. J., Mattanovich, D. & Gasser, B. The impact of ERAD on recombinant protein secretion in Pichia pastoris (syn Komagataella spp.). *Microbiology* **164**, 453–463 (2018).
22. Akopian, D., Shen, K., Zhang, X. & Shan, S.-O. Signal recognition particle: an essential protein-targeting machine. *Annu. Rev. Biochem.* **82**, 693–721 (2013).
23. Nyathi, Y., Wilkinson, B. M. & Pool, M. R. Co-translational targeting and translocation of proteins to the endoplasmic reticulum. *Biochim. Biophys. Acta* **1833**, 2392–2402 (2013).
24. Alva, T. R., Riera, M. & Chartron, J. W. Translational landscape and protein biogenesis demands of the early secretory pathway in Komagataella phaffii. *Microb. Cell Fact.* **20**, 19 (2021).
25. *The Unfolded Protein Response and Cellular Stress, Part C*. (Academic Press, 2011).
26. Ingolia, N. T. Ribosome Footprint Profiling of Translation throughout the Genome. *Cell* **165**, 22–33 (2016).
27. de Bruijn, F. J. *Stress and Environmental Regulation of Gene Expression and Adaptation in Bacteria*. (John Wiley & Sons, 2016).
28. Taggart, J. C. & Li, G.-W. Production of Protein-Complex Components Is Stoichiometric and Lacks General Feedback Regulation in Eukaryotes. *Cell Syst* **7**, 580–589.e4 (2018).
29. Li, G.-W., Burkhardt, D., Gross, C. & Weissman, J. S. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell* **157**, 624–635 (2014).
30. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534–1550 (2012).
31. Chung, B. Y. *et al.* The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for Ribo-seq data analysis. *RNA* **21**, 1731–1745 (2015).
32. Faridani, O. R. *et al.* Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.* **34**, 1264–1266 (2016).
33. Archer, S. K., Shirokikh, N. E. & Preiss, T. Selective and flexible depletion of problematic sequences from RNA-seq libraries at the cDNA stage. *BMC Genomics* **15**, 401 (2014).
34. Benes, V., Blake, J. & Doyle, K. Ribo-Zero Gold Kit: improved RNA-seq results after removal of cytoplasmic and mitochondrial ribosomal RNA. *Nat. Methods* **8**, iii–iv (2011).
35. McGlincy, N. J. & Ingolia, N. T. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112–129 (2017).
36. Becker, A. H., Oh, E., Weissman, J. S., Kramer, G. & Bukau, B. Selective ribosome profiling as a tool for studying the interaction of chaperones and targeting factors with nascent polypeptide chains and ribosomes. *Nat. Protoc.* **8**, 2212–2239 (2013).
37. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
38. Vogl, T. *et al.* A Toolbox of Diverse Promoters Related to Methanol Utilization: Functionally Verified Parts for Heterologous Pathway Expression in Pichia pastoris. *ACS Synth. Biol.* **5**, 172–186 (2016).
39. Yano, T., Takigami, E., Yurimoto, H. & Sakai, Y. Yap1-regulated glutathione redox system

curtails accumulation of formaldehyde and reactive oxygen species in methanol metabolism of Pichia pastoris. *Eukaryot. Cell* **8**, 540–549 (2009).

40. Cámara, E. *et al.* Increased dosage of AOX1 promoter-regulated expression cassettes leads to transcription attenuation of the methanol metabolism in Pichia pastoris. *Sci. Rep.* **7**, 44302 (2017).
41. Zhang, H. *et al.* Alcohol oxidase (AOX1) from Pichia pastoris is a novel inhibitor of prion propagation and a potential ATPase. *Mol. Microbiol.* **71**, 702–716 (2009).
42. de Hoop, M. J. *et al.* Overexpression of alcohol oxidase in Pichia pastoris. *FEBS Lett.* **291**, 299–302 (1991).
43. Küberl, A. *et al.* High-quality genome sequence of Pichia pastoris CBS7435. *J. Biotechnol.* **154**, 312–320 (2011).
44. Krainer, F. W. *et al.* Recombinant protein expression in Pichia pastoris strains with an engineered methanol utilization pathway. *Microb. Cell Fact.* **11**, 22 (2012).
45. Vanz, A. L., Nimtz, M. & Rinas, U. Decrease of UPR- and ERAD-related proteins in Pichia pastoris during methanol-induced secretory insulin precursor production in controlled fed-batch cultures. *Microb. Cell Fact.* **13**, 23 (2014).
46. Vanz, A. L. *et al.* Physiological response of Pichia pastoris GS115 to methanol-induced high level production of the Hepatitis B surface antigen: catabolic adaptation, stress responses, and autophagic processes. *Microb. Cell Fact.* **11**, 103 (2012).
47. Roth, G., Vanz, A. L., Lünsdorf, H., Nimtz, M. & Rinas, U. Fate of the UPR marker protein Kar2/Bip and autophagic processes in fed-batch cultures of secretory insulin precursor producing Pichia pastoris. *Microb. Cell Fact.* **17**, 123 (2018).
48. Liebermeister, W. *et al.* Visual account of protein investment in cellular functions. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8488–8493 (2014).
49. Otto, A. *et al.* Systems-wide temporal proteomic profiling in glucose-starved Bacillus subtilis. *Nat. Commun.* **1**, 137 (2010).
50. Bernhardt, J., Funke, S., Hecker, M. & Siebourg, J. Visualizing Gene Expression Data via Voronoi Treemaps. in *2009 Sixth International Symposium on Voronoi Diagrams* (IEEE, 2009). doi:10.1109/isvd.2009.33.
51. Tafrishi, A. *et al.* Functional genomic screening in Komagataella phaffii enabled by high-activity CRISPR-Cas9 library. *Metab. Eng.* **85**, 73–83 (2024).
52. Kroukamp, H. *et al.* Strain Breeding Enhanced Heterologous Cellobiohydrolase Secretion by Saccharomyces cerevisiae in a Protein Specific Manner. *Biotechnol. J.* **12**, (2017).
53. Gasser, B., Sauer, M., Maurer, M., Stadlmayr, G. & Mattanovich, D. Transcriptomics-based identification of novel factors enhancing heterologous protein secretion in yeasts. *Appl. Environ. Microbiol.* **73**, 6499–6507 (2007).
54. Gasser, B. *et al.* Monitoring of transcriptional regulation in Pichia pastoris under protein production conditions. *BMC Genomics* **8**, 179 (2007).
55. Kang, Z., Huang, H., Zhang, Y., Du, G. & Chen, J. Recent advances of molecular toolbox construction expand Pichia pastoris in synthetic biology applications. *World J. Microbiol. Biotechnol.* **33**, 19 (2017).
56. Zhou, Y., Raju, R., Alves, C. & Gilbert, A. Debottlenecking protein secretion and reducing protein aggregation in the cellular host. *Curr. Opin. Biotechnol.* **53**, 151–157 (2018).
57. Vogl, T. & Glieder, A. Regulation of Pichia pastoris promoters and its consequences for protein production. *N. Biotechnol.* **30**, 385–404 (2013).
58. Kallehauge, T. B. *et al.* Ribosome profiling-guided depletion of an mRNA increases cell growth rate and protein secretion. *Sci. Rep.* **7**, 40388 (2017).
59. Kol, S. *et al.* Multiplex secretome engineering enhances recombinant protein production and

purity. *Nat. Commun.* **11**, 1908 (2020).

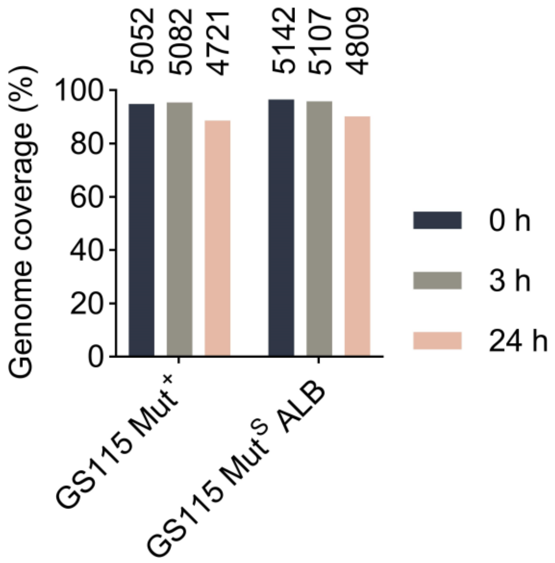
60. Fischer, J. E. & Glieder, A. Current advances in engineering tools for Pichia pastoris. *Curr. Opin. Biotechnol.* **59**, 175–181 (2019).
61. Radoman, B. *et al.* The Degree and Length of O-Glycosylation of Recombinant Proteins Produced in Pichia pastoris Depends on the Nature of the Protein and the Process Type. *Biotechnol. J.* **16**, e2000266 (2021).
62. Delic, M. *et al.* Overexpression of the transcription factor Yap1 modifies intracellular redox conditions and enhances recombinant protein secretion. *Microb. Cell Fact.* **1**, 376–386 (2014).
63. Lamour, J., Wan, C., Zhang, M., Zhao, X. & Den Haan, R. Overexpression of endogenous stress-tolerance related genes in Saccharomyces cerevisiae improved strain robustness and production of heterologous cellobiohydrolase. *FEMS Yeast Res.* **19**, (2019).
64. Harding, H. P. *et al.* An integrated stress response regulates amino acid metabolism and resistance to oxidative stress. *Mol. Cell* **11**, 619–633 (2003).
65. Morotti, M. *et al.* Increased expression of glutamine transporter SNAT2/SLC38A2 promotes glutamine dependence and oxidative stress resistance, and is associated with worse prognosis in triple-negative breast cancer. *Br. J. Cancer* **124**, 494–505 (2021).
66. Yano, T., Yurimoto, H. & Sakai, Y. Activation of the oxidative stress regulator PpYap1 through conserved cysteine residues during methanol metabolism in the yeast Pichia pastoris. *Biosci. Biotechnol. Biochem.* **73**, 1404–1411 (2009).
67. Blevins, W. R. *et al.* Extensive post-transcriptional buffering of gene expression in the response to oxidative stress in baker's yeast. *bioRxiv* 501478 (2019) doi:10.1101/501478.
68. Gerashchenko, M. V., Lobanov, A. V. & Gladyshev, V. N. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17394–17399 (2012).
69. Shenton, D. *et al.* Global translational responses to oxidative stress impact upon multiple levels of protein synthesis. *J. Biol. Chem.* **281**, 29011–29021 (2006).
70. Vogel, C., Silva, G. M. & Marcotte, E. M. Protein expression regulation under oxidative stress. *Mol. Cell. Proteomics* **10**, M111.009217 (2011).
71. Kastberg, L. L. B., Ard, R., Jensen, M. K. & Workman, C. T. Burden Imposed by Heterologous Protein Production in Two Major Industrial Yeast Cell Factories: Identifying Sources and Mitigation Strategies. *Front Fungal Biol* **3**, 827704 (2022).
72. Niesen, M. J. M., Zimmer, M. H. & Miller, T. F., 3rd. Dynamics of Co-translational Membrane Protein Integration and Translocation via the Sec Translocon. *J. Am. Chem. Soc.* **142**, 5449–5460 (2020).
73. Ng, D. T., Brown, J. D. & Walter, P. Signal sequences specify the targeting route to the endoplasmic reticulum membrane. *J. Cell Biol.* **134**, 269–278 (1996).
74. Ast, T., Cohen, G. & Schuldiner, M. A network of cytosolic factors targets SRP-independent proteins to the endoplasmic reticulum. *Cell* **152**, 1134–1145 (2013).
75. Berndt, U., Oellerer, S., Zhang, Y., Johnson, A. E. & Rospert, S. A signal-anchor sequence stimulates signal recognition particle binding to ribosomes from inside the exit tunnel. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 1398–1403 (2009).
76. Berkovits, B. D. & Mayr, C. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**, 363–367 (2015).
77. Chartron, J. W., Hunt, K. C. L. & Frydman, J. Cotranslational signal-independent SRP preloading during membrane targeting. *Nature* **536**, 224–228 (2016).
78. Rapoport, T. A. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature* **450**, 663–669 (2007).

79. Shao, S. & Hegde, R. S. A calmodulin-dependent translocation pathway for small secretory proteins. *Cell* **147**, 1576–1588 (2011).
80. Burgard, J. *et al.* The secretome of Pichia pastoris in fed-batch cultivations is largely independent of the carbon source but changes quantitatively over cultivation time. *Microb. Biotechnol.* **13**, 479–494 (2020).
81. Klis, F. M., Mol, P., Hellingwerf, K. & Brul, S. Dynamics of cell wall structure in Saccharomyces cerevisiae. *FEMS Microbiol. Rev.* **26**, 239–256 (2002).
82. Mrsa, V. *et al.* Deletion of new covalently linked cell wall glycoproteins alters the electrophoretic mobility of phosphorylated wall components of Saccharomyces cerevisiae. *J. Bacteriol.* **181**, 3076–3086 (1999).
83. Shimoi, H., Kitagaki, H., Ohmori, H., Iimura, Y. & Ito, K. Sed1p is a major cell wall protein of Saccharomyces cerevisiae in the stationary phase and is involved in lytic enzyme resistance. *J. Bacteriol.* **180**, 3381–3387 (1998).
84. van der Vaart, J. M., Caro, L. H., Chapman, J. W., Klis, F. M. & Verrips, C. T. Identification of three mannoproteins in the cell wall of Saccharomyces cerevisiae. *J. Bacteriol.* **177**, 3104–3110 (1995).
85. Thorpe, G. W., Fong, C. S., Alic, N., Higgins, V. J. & Dawes, I. W. Cells have distinct mechanisms to maintain protection against different reactive oxygen species: oxidative-stress-response genes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 6564–6569 (2004).
86. Kumita, J. R. *et al.* Impact of the native-state stability of human lysozyme variants on protein secretion by Pichia pastoris. *FEBS J.* **273**, 711–720 (2006).
87. Hesketh, A. R., Castrillo, J. I., Sawyer, T., Archer, D. B. & Oliver, S. G. Investigating the physiological response of Pichia (Komagataella) pastoris GS115 to the heterologous expression of misfolded proteins using chemostat cultures. *Appl. Microbiol. Biotechnol.* **97**, 9747–9762 (2013).
88. Coen, M. L., Lerner, C. G., Capobianco, J. O. & Goldman, R. C. Synthesis of yeast cell wall glucan and evidence for glucan metabolism in a Saccharomyces cerevisiae whole cell system. *Microbiology* **140 (Pt 9)**, 2229–2237 (1994).
89. Li, J. *et al.* Improved cellulase production in recombinant Saccharomyces cerevisiae by disrupting the cell wall protein-encoding gene CWP2. *J. Biosci. Bioeng.* **129**, 165–171 (2020).
90. Zhang, M. *et al.* Deletion of yeast CWP genes enhances cell permeability to genotoxic agents. *Toxicol. Sci.* **103**, 68–76 (2008).
91. Li, J. *et al.* Increasing extracellular cellulase activity of the recombinant Saccharomyces cerevisiae by engineering cell wall-related proteins for improved consolidated processing of carbon neutral lignocellulosic biomass. *Bioresour. Technol.* **365**, 128132 (2022).
92. Chen, N. *et al.* Systematic genetic modifications of cell wall biosynthesis enhanced the secretion and surface-display of polysaccharide degrading enzymes in Saccharomyces cerevisiae. *Metab. Eng.* **77**, 273–282 (2023).
93. Archer, S. K., Shirokikh, N. E. & Preiss, T. Probe-Directed Degradation (PDD) for Flexible Removal of Unwanted cDNA Sequences from RNA-Seq Libraries. *Curr. Protoc. Hum. Genet.* **85**, 11.15.1–11.15.36 (2015).
94. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
95. Jain, M. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* **12**, 351–356 (2015).
96. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

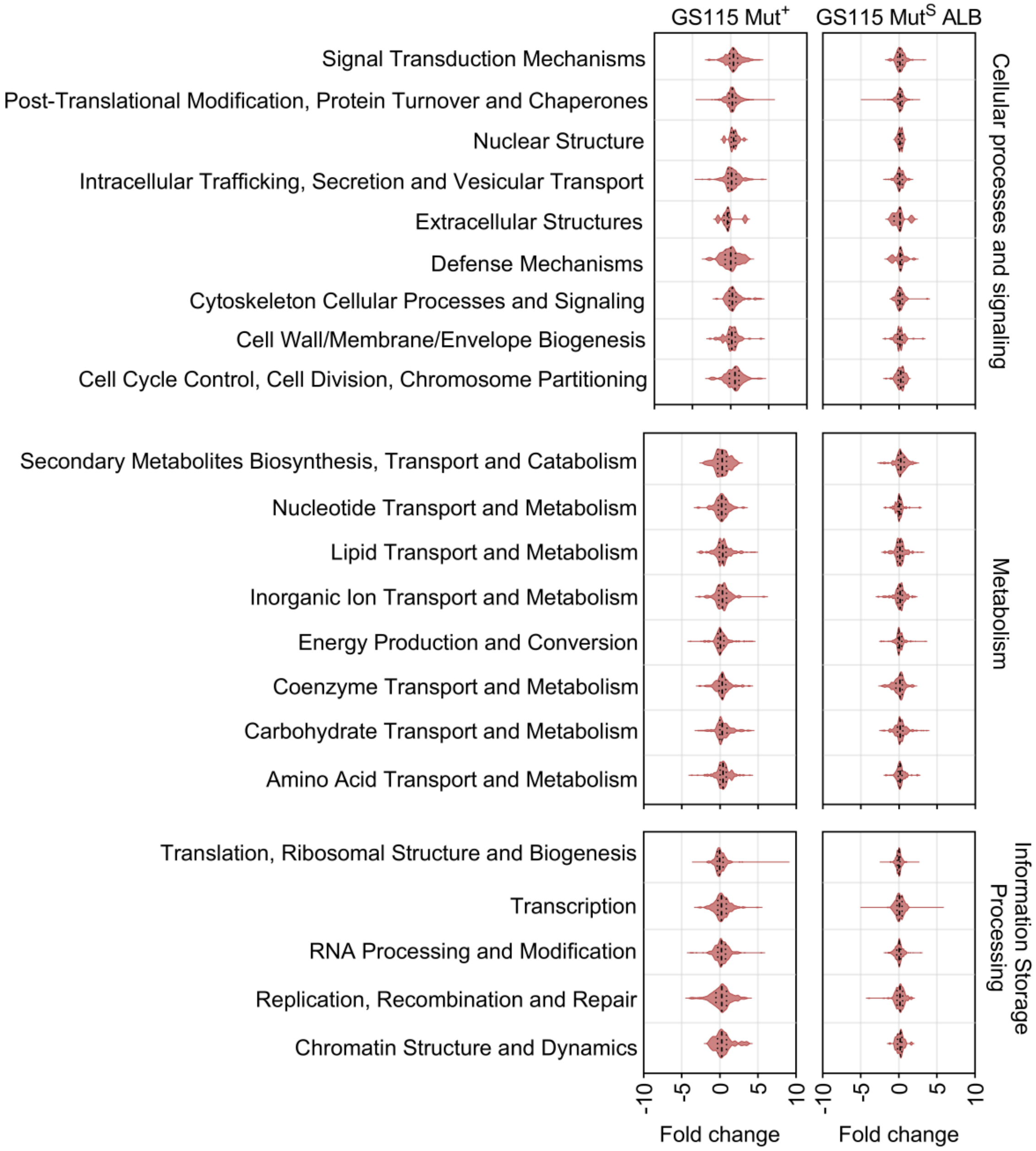
97. Love, K. R. *et al.* Comparative genomics and transcriptomics of Pichia pastoris. *BMC Genomics* **17**, (2016).
98. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
99. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
100. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
101. Popa, A. *et al.* RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing. *F1000Res.* **5**, 1309 (2016).
102. Mohammad, F., Green, R. & Buskirk, A. R. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *Elife* **8**, (2019).
103. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–93 (2016).
104. Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H. & Winther, O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**, 3387–3395 (2017).
105. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).
106. Tsirigos, K. D., Peters, C., Shu, N., Käll, L. & Elofsson, A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.* **43**, W401–7 (2015).
107. Pierleoni, A., Martelli, P. L. & Casadio, R. PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* **9**, 392 (2008).
108. Dalvie, N. C. *et al.* Host-Informed Expression of CRISPR Guide RNA for Genomic Engineering in Komagataella phaffii. *ACS Synth. Biol.* **9**, 26–35 (2020).
109. Schwartz, C. & Wheeldon, I. CRISPR-Cas9-Mediated Genome Editing and Transcriptional Control in Yarrowia lipolytica. *Methods Mol. Biol.* **1772**, 327–345 (2018).

4.6 Supplementary information

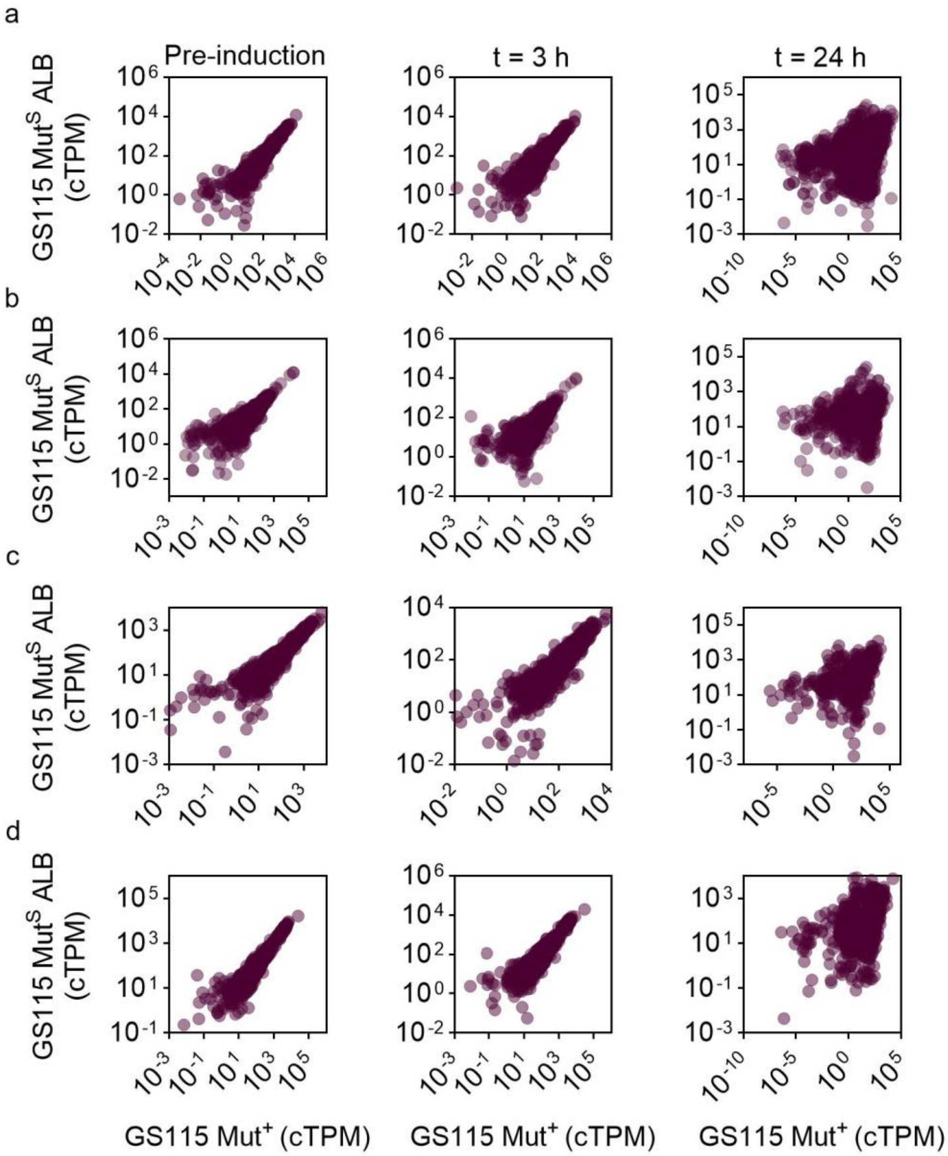
4.6.1 Supplementary figures



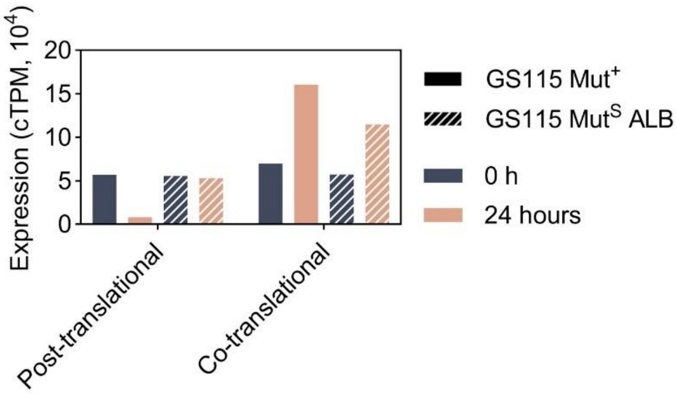
Supplementary Fig. 4- 1. Genome-wide coverage of gene expression obtained from Ribo-seq data. The numbers on each bar represent the number of detected genes in the final data.



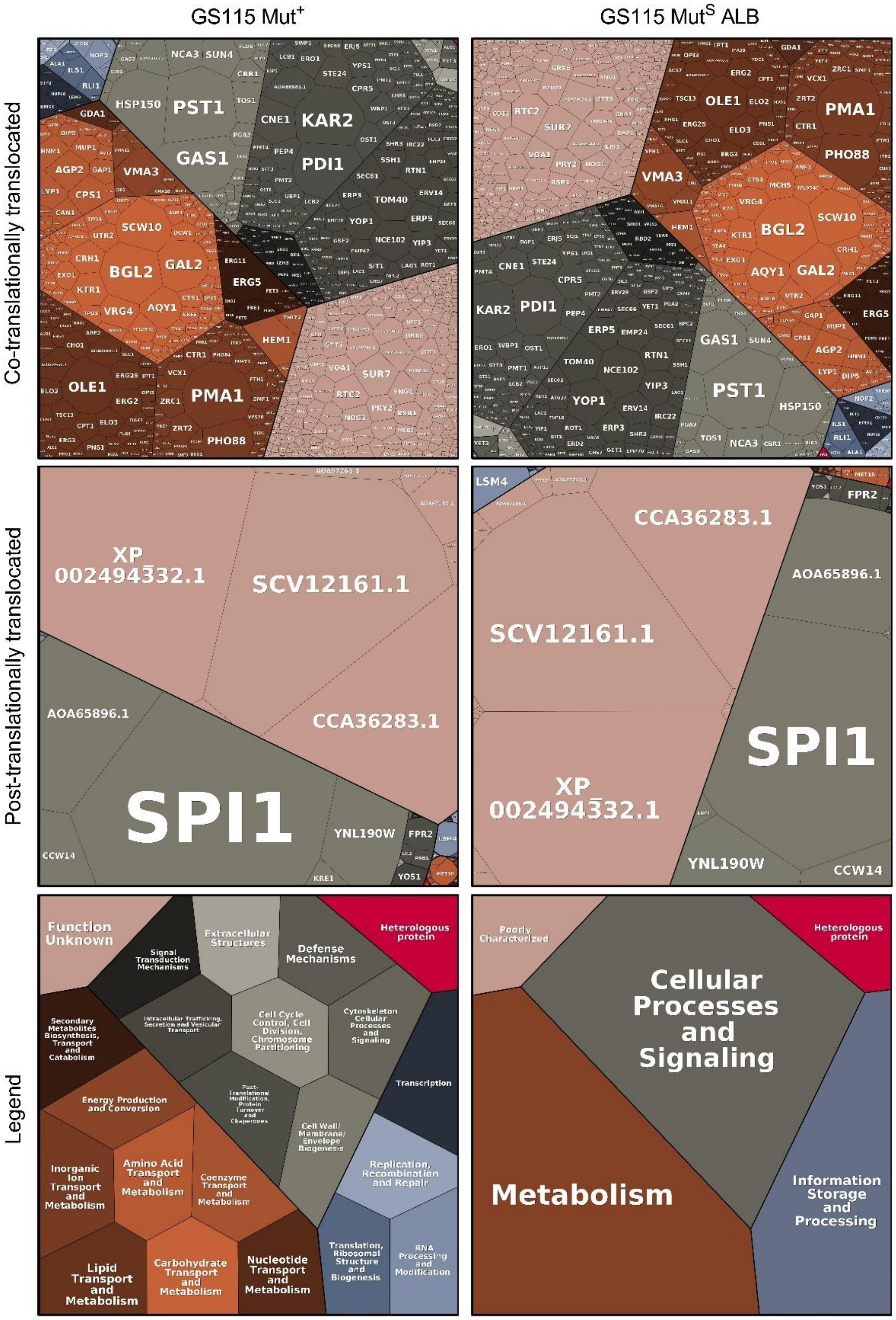
Supplementary Fig. 4- 2. Fold change production of nascent chains belonging to different ontological categories. Fold change is compared in GS115 Mut$^+$ and GS115 Mut^S ALB 3-hours after induction in buffered methanol media (BMMY). Fold change is calculated as the log$_2$ ratio of expressed genes three hours after methanol induction compared to the expression levels before induction.



Supplementary Fig. 4- 3. Divergence of translational landscape after heterologous expression for ontological categories. **a)** Cell processes and signaling, **b)** Poorly characterized, **c)** Metabolism, and **d)** Information storage and processing.



Supplementary Fig. 4- 4. Total nascent chain production translocating into the ER co- and post-translationally. 56 and 931 protein products are predicted to enter the ER post- and co-translationally.


GS115 Mut+
GS115 MutS ALB
Co-translationally translocated
Post-translationally translocated
Legend
NOP2
ILS1
RLI1
GDA1
HSP150
PST1
NCA3
SUN4
GAS1
KAR2
CNE1
PDI1
STE24
ERO1
YPS1
CPR5
OST1
SSH1
RTN1
SEC61
TOM40
YOP1
ERP5
NCE102
ERV14
YIP3
AGP2
VMA3
CPS1
SCW10
BGL2
GAL2
CRH1
KTR1
VRG4
AQY1
ERG11
ERG5
HEM1
OLE1
PMA1
PHO88
ERG2
ZRC1
CTR1
VCX1
RTC2
SUR7
VOA1
PRY2
EMP24
ERP3
GRX6
PMT4
SEC66
LSM4
MET10
XP 002494332.1
SCV12161.1
CCA36283.1
AOA65896.1
SPI1
YNL190W
CCW14
FPR2
KRE1
YOS1
Function Unknown
Signal Transduction Mechanisms
Extracellular Structures
Defense Mechanisms
Heterologous protein
Secondary Metabolites Biosynthesis, Transport and Catabolism
Intracellular Trafficking, Secretion and Vesicular Transport
Cell Cycle Control, Cell Division, Chromosome Partitioning
Cytoskeleton Cellular Processes and Signaling
Transcription
Energy Production and Conversion
Post-Translational Modification, Protein Turnover and Chaperones
Cell Wall/ Membrane/ Envelope Biogenesis
Inorganic Ion Transport and Metabolism
Amino Acid Transport and Metabolism
Coenzyme Transport and Metabolism
Replication, Recombination and Repair
Lipid Transport and Metabolism
Carbohydrate Transport and Metabolism
Nucleotide Transport and Metabolism
Translation, Ribosomal Structure and Biogenesis
RNA Processing and Modification
Poorly Characterized
Cellular Processes and Signaling
Metabolism
Information Storage and Processing

Supplementary Fig. 4- 5. Co and post-translational flux through the ER for GS115 Mut$^+$ and GS115 Mut^S ALB strains pre-induction. Non-mitochondrial proteins are predicted to enter the secretory pathway co-translationally if they have greater than log$_2$ membrane enrichment in YPD studies. Gene products are grouped by ontological function using COG scores predicted by EggNOG v5.0. Cell sizes are calculated using cTPM scores and represent relative quantities of nascent chains produced per gene. Tessellation plots are made using www.bionic-vis.biologie.uni-greifswald.de [1–3]

4.6.2 Supplementary tables

Supplementary Table 4- 1. Oligos designed for probe-directed degradation.

Probe sequences [a]	Read abundance [b]
GTTGGTGCGTCTACGCATCTCCGAC	10,400,000
CCGTGGGTGAGACGGTCCTAAGGGC	1,400,000
CATACCCGTGAAAATTTGGTTTATT	1,000,000
TGTTATTCCCCCGCCCGTACTGACA	1,000,000
CAAAGAGGGTGATAGCCCCGTGGCA	760,000
CCTCCGCCCATTCTCAAACTTTAAA	600,000
AGGGCAGTAAAACCCGAAGAGCGTG	500,000
CAAAGAGGGTGATAGCCCCGTAGCA	450,000
TGTGTGGCGAAGACCTGCTTTAGTG	400,000
GAGTGTTCAAGGCAGTAGTTGAATA	300,000
ATACAGGGAGGGTGGGGTGAGT	300,000
CTAGACCCCCTCAGTGGGCCATTTT	300,000
GTTTAGTTCCATGAGGTAAAGCAAT	170,000
CGCCAAGGACGTTTTCATTAATCAA	165,000
ACTCTGGTGGAGGCCCGCAGCGGTT	130,000
TTATCGACCAACCCAGAACTG	95,000
CCATATCTAGCAGAAAGCACCGTTT	86,084
AACGGCGGGAGTAACTATGACTCT	75,000
AGAAACCTCCAGGCGGGGAGTTTGG	70,000
ATCGTTGCGAGAGCCAAGAGATCCG	566

[a] Complementary oligonucleotides to Ribo-seq sequences mapped most highly to GS115 rRNA

[b] Ribo-seq reads aligned to GS115 rRNA

Supplementary Table 4- 2. Strains used in this study.

Strain	Genotype	Reference
E. coli TOP10	F- mcrA Δ(mrr-hsdRMS-mcrBC) φ80lacZΔM15 ΔlacX74 recA1 araD139 Δ(ara-leu) 7697 galU galK rpsL (StrR) endA1 nupG λ	Thermo Fisher Scientific
GS115 Mut^{+}	*his4*	Invitrogen
GS115 Mut^{S} ALB	*his4, aox1::HSA*	Invitrogen
GS115 Mut^{S} ALB *gal2*	*his4, gal2, aox1::HSA*	This study
GS115 Mut^{S} ALB *aoa65896.1*	*his4, aoa65896.1, aox1::HSA*	This study
GS115 Mut^{S} ALB *bgl2*	*his4, bgl2, aox1::HSA*	This study
GS115 Mut^{S} ALB *scv12161.1*	*his4, scv12161.1, aox1::HSA*	This study
GS115 Mut^{S} ALB *ydr134c*	*his4, ydr134c, aox1::HSA*	This study
GS115 Mut^{S} ALB *gal2 ydr134c*	*his4, gal2, ydr134c***,** *aox1::HSA*	This study
GS115 Mut^{S} ALB *bgl2 ydr134c*	*his4, bgl2***,** *ydr134c, aox1::HSA*	This study
GS115 Mut^{S} ALB *bgl2 gal2*	*his4, gal2, bgl2, aox1::HSA*	This study
GS115 Mut^{S} ALB *scv12161.1 aoa65896.1*	*his4, scv12161.1, aoa65896.1, aox1::HSA*	This study
GS115 Mut^{S} ALB *gal2 aoa65896.1*	*his4, gal2, aoa65896.1, aox1::HSA*	This study
GS115 Mut^{S} ALB *ydr134c aoa65896.1*	*his4, ydr134c, aoa65896.1 , aox1::HSA*	This study
GS115 Mut^{S} ALB *scv12161.1 gal2*	*his4, scv12161.1, gal2, aox1::HSA*	This study
GS115 Mut^{S} ALB *scv12161.1 ydr134c*	*his4, scv12161.1, ydr134c ,aox1::HSA*	This study
GS115 Mut^{S} ALB *gal2 aoa65896.1 ydr134c*	*his4, gal2, aoa65896.1, ydr134c, aox1::HSA*	This study
GS115 Mut^{S} ALB *gal2 aoa65896.1 ydr134c*	*his4, gal2, aoa65896.1, ydr134c, aox1::HSA*	This study
GS115 Mut^{S} ALB *gal2 bgl2 ydr134c*	*his4, gal2, bgl2, ydr134c, aox1::HSA*	This study
GS115 Mut^{S} ALB *gal2 aoa65896.1 ydr134c scv12161.1*	*his4, gal2, aoa65896.1, ydr134c, scv12161.1, aox1::HSA*	This study
GS115 Mut^{S} ALB *gal2 aoa65896.1 ydr134c bgl2*	*his4, gal2, aoa65896.1, ydr134c, bgl2, aox1::HSA*	This study
GS115 Mut^{S} ALB *gal2*	*his4, gal2, aoa65896.1, ydr134c,*	This study

aoa65896.1 ydr134c bgl2 scv12161.1	*bgl2, scv12161.1 , aox1::HSA*

Supplementary Table 4- 3. sgRNAs used in this study.

sgRNA target	sgRNA sequence
AOA65896.1	ACAAGAGGTGATAGTCAGAA
YDR134C	GAACTCTCCAGAATCAGCAA
GAL2	GACATCCCAGTCAAACCCAA
BGL2	GGAATAAGCTCTAATAGCAA
SCV12161.1	TTCGTTTTGAGCTTGCACAA

Supplementary Table 4- 4. Primers used in this study.

Primer Name	Primer sequence
AOA65896.1.FOR	TTCCTCAACTCACTGTTTCAGTTTATTCCAAC
AOA65896.1.REV	GTGAGAGCTGGTCTTAGCTGGAG
BGL2.FOR	ATCTGAAGCTGGCAAGTCGTC
BGL2.REV	GATCTTTAATCTTAAAACACTGGCTGCG
GAL2.FOR	TAATATGAGTTCAACAGATATCCAAGGTGATCAAG
GAL2.REV	AAGGTAATACGTTTCACCGTTAAACTGT
SCV12161.1.FOR	CCACAAAATTTCAGCGAGCAACAG
SCV12161.1.REV	AGTCCTCACCTACAGCCAACT
YDR134C.FOR	ATAATGCTAACCAAGGTTATTTCACTCGC
YDR134C.REV	AGTGTAAGAAACACATTCGGGGT

References

1. Liebermeister, W. *et al.* Visual account of protein investment in cellular functions. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8488–8493 (2014).
2. Otto, A. *et al.* Systems-wide temporal proteomic profiling in glucose-starved Bacillus subtilis. *Nat. Commun.* **1**, 137 (2010).
3. Bernhardt, J., Funke, S., Hecker, M. & Siebourg, J. Visualizing Gene Expression Data via Voronoi Treemaps. in *2009 Sixth International Symposium on Voronoi Diagrams* (IEEE, 2009). doi:10.1109/isvd.2009.33.

Chapter 5: Conclusion

Komagataella phaffii has been widely used in the biotechnology industry, because of its ability to produce and secrete high titers of recombinant proteins, its amenability to genetic engineering, its capability to perform post-translational modifications, and the presence of synthetic biology tools. However, despite its widespread use and proven advantages, there is still a significant need for more advanced genetic engineering tools to perform functional genetic screens in a high-throughput manner and to fully harness its potential.

This dissertation addresses the challenge of limited genome engineering tools in non-model organisms by expanding the available toolkit for *K. phaffii*, thereby accelerating design-build-test-learn cycles for its metabolic engineering. Pooled CRISPR screens are powerful tools to investigate phenotype-to-genotype associations and identify essential genes under a specific condition, but their use in non-model organisms has been constrained by difficulties in predicting and designing highly active sgRNAs, which are essential for accurate screening. To overcome this, we provided a detailed design protocol for a highly active, genome-wide knockout sgRNA library in Chapter 2, that enables the study of all genes in *K. phaffii's* genome. This protocol ensures sgRNAs are unique to their target sites, predicted to be active, and unlikely to form secondary RNA structures. In Chapter 3, by disrupting the NHEJ DNA repair pathway and performing growth screens with the 6-fold library, we measured the activity of each sgRNA, demonstrating that this library design approach can be applied to generate highly active sgRNAs for any species.

While we demonstrated the application of this design protocol to create an active genome-wide sgRNA knockout library in *K. phaffii*, there are additional promising avenues to explore. Many genes in the genome are essential for growth, meaning that knocking them out would result in cell death or extremely unhealthy growth. To study the impact of these essential genes as well as the influence of up- or down-regulation of other non-essential genes on specific phenotypes, CRISPRi and CRISPRa systems can be used to downregulate or upregulate gene expression, respectively. The protocol outlined in Chapter 2 can also be applied to design highly specific sgRNAs that target promoter regions, allowing for precise control of gene expression. Additionally, by transforming the library into NHEJ-deficient cells, the activity profiles of any library can be accurately measured, further increasing our confidence in identifying hits to improve any phenotype. In Chapter 3, we utilized the highly active knockout library to establish a comprehensive consensus set of essential genes for *K. phaffii*. This consensus set was defined as the union of essential genes identified through our CRISPR screen combined with those from a previous transposon-insertion study. The number of essential genes we predicted aligns with expectations based on analyses of other yeasts, such as *S. cerevisiae*, *Schizosaccharomyces pombe*, *Yarrowia lipolytica*, and *Kluyveromyces marxianus*, further validating our findings. Gene Ontology (GO) enrichment analysis also provided additional validation by highlighting significant enrichment in critical biological processes, including transcription, translation, cell cycle regulation, and ribosome biogenesis—processes fundamental to cellular function and survival.

Moreover, by comparing *K. phaffii* to other yeast species, we identified a core set of conserved essential genes that play crucial roles in vital cellular functions. Beyond this core set, we uncovered *K. phaffii*-exclusive essential genes, which GO and KEGG enrichment analyses revealed to be predominantly involved in protein localization, transport, secretion, and the N-glycosylation pathway. Notably, many of these essential genes have been previously studied as overexpression targets to enhance secretion capacity or improve glycosylation, demonstrating their significance in improving these phenotypes. Consequently, this chapter not only identifies a list of essential genes within these pathways as candidates for further research but also opens new avenues for optimizing secretion processes in this non-conventional yeast. Furthermore, this highly active library can be leveraged to study and uncover additional genes involved in specific pathways, including methanol assimilation, or to identify knockouts that could enhance cell survival under various stress conditions. This powerful tool provides a valuable resource for future research aimed at engineering *K. phaffii* for improved performance in industrial applications.

As discussed earlier, one of the key characteristics of *K. phaffii* that has led to its extensive use in the industry is its ability to produce and secrete high levels of recombinant products. The success of this process is significantly influenced by the proper functioning of the secretory pathway and the correct folding of proteins. The first step in the secretion pathway is the translocation of nascent chain polypeptides to the ER membrane which involves with the coordination between various cell machinery including translocon pores, chaperones, and other ER-resident proteins. In Chapter 4, we utilized Ribo-seq and next generation sequencing to gain deeper insights into host protein synthesis and endoplasmic reticulum trafficking of *K. phaffii* when grown under heterologous expression conditions. We focused on comparing the translatomes of two commonly used industrial strains, GS115 Mut$^+$ and GS115 Mut^S ALB, before and after methanol induction. We enhanced our Ribo-seq pipeline by implementing more efficient rRNA depletion methods and developing a technique that minimizes the need for biological replicates to accurately determine read count thresholds for differential expression analysis.

Our analysis revealed that the slower-growing Mut^S strain, despite its lower methanol utilization due to the activity of only the *AOX2* gene, achieves higher heterologous protein yields than the faster-growing Mut$^+$ strain. This unexpected result led us to investigate further, and we found that the Mut$^+$ strain experiences higher oxidative stress due to producing more toxic compounds such as formaldehyde, which negatively impacts its protein production capabilities. Through Ribo-seq, we detected significant markers of oxidative stress in the Mut$^+$ strain, such as translation initiation discrepancies, slower elongation rates, and increased expression of stress response genes. These findings suggest that the oxidative stress in the Mut$^+$ strain could be a key factor limiting its productivity. One way to further continue engineering *K. phaffii* for higher heterologous protein production based on our findings is to overexpress genes involved in oxidative stress management to alleviate the stress caused by methanol metabolism.

Additionally, our investigation into endoplasmic reticulum (ER) trafficking revealed that the ER flux was greater in the Mut^S strain, with distinct patterns of protein translocation before and after methanol induction. This differential ER activity highlights potential targets for rational strain engineering, such as modifying the expression of cell wall components, to further enhance protein secretion. From our findings, we identified five highly expressed, non-essential genes predicted to enter the early secretory pathway. Using the CRISPR-Cas9 system, we knocked out these genes either individually or in combination. Remarkably, we discovered that a combination of three gene knockouts led to a 35% improvement in human serum albumin secretion levels. The methodologies and insights presented in this chapter not only provide a framework for enhancing heterologous protein production in *K. phaffii* but also serve as a guide for applying these strategies to other industrially relevant, yet less understood, organisms.