UNIVERSITY OF CALIFORNIA
RIVERSIDE

Rational Designs for Increased Bioproduction in *Komagataella phaffii*

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

Doctor of Philosophy

in

Bioengineering

by

Troy Alva

September 2021

Dissertation Committee:

Dr. Ian Wheeldon, Chairperson
Dr. Joshua T. Morgan
Dr. Thomas Girke

The Dissertation of Troy Alva is approved:

Committee Chairperson

University of California, Riverside

Acknowledgements

Much of science is a collaborative enterprise, and my contributions could not have been without the many brilliant and devoted scientists I have had the great pleasure of working with. For these people, I express my sincerest gratitude.

Thank you to my doctoral advisor Dr. Ian Wheeldon for your mentorship, guidance, decisiveness, communication, and trust. I appreciate all that you have done for me and the many lessons you have taught me. You taught me that even the most brilliant can be humble, that clear communication is effective communication, and to shoot high and not limit the impact of my work. I am grateful for your time and investment.

Thank you to my master's advisor, Dr. Victor G.J. Rodgers for providing me the opportunity to call myself an engineer. You taught me to fiercely defend my work and that, to succeed, you need to "dream" of solutions to the problems you face.

Thank you to my former doctoral advisor, Dr. Justin W. Chartron for your patience, kindness, and support. I will remember our scientific discussions, obsessions over data together, and interrogations of what could be. You taught me the value of self learning, to treat my colleagues with respect, and to give myself grace when things become challenging.

Thank you to my committee, Dr. Joshua T. Morgan and Dr. Thomas Girke, for your guidance, support, and for challenging me to keep pushing the boundaries. Josh, it was a pleasure to have the opportunity to collaborate with you on multiple projects. I appreciate your guidance as a mentor and graduate student advisor.

Thank you to the graduate and post-doctoral colleagues I have had the pleasure to know and work with. I treasure your friendship and look forward to the impact you all will undoubtedly make on the world. While I cannot name you all, I would like to express my gratitude to the following: Dr. Christopher Hale,

Dr. Dieanira Erudaitius, Raymond Yeung, Jennifer Yang, Heran Bhakta, Dr. Nehamiah Zewde, Alexander Brunelle, Brian Lupish, Mario Leon Lopez, Dominic Biondo, Aida Tafrishi, Dr. Clifford Morrison, Dr. Xuye Lang, Adithya Ramesh, Varun Trivedi, Ben Rammelsberg, Jordan Dagan, Mengwan Li, Nick Robertson, Sarah Thorwall, Shuang Wei, Wenguang Wang, Xiao Hong, Xiyan Sun, Shane Kennedy, Joseph Schwan, Wayne Leu, Adam Witmer, and James Stumpff.

Thank you to the army of undergraduates I had the pleasure of working with. You all taught me accountability and helped me realize how much I enjoy leading.

Finally, thank you to the staff at UC Riverside for helping create a positive environment that allowed me to grow. Thank you Hong Xu for providing me with support throughout my time at UC Riverside. Thank you Dr. B. Hyle Park for your guidance as my former graduate student advisor and for your help while I led the bioengineering mentorship program. Thank you to all of the bioengineering staff for your guidance, tutelage, and for sharpening my skills as an engineer. Thank you Martin Kleckner and Art Salyer at UC Riverside's office of technology partnerships for your guidance and insights for applying my skills in entrepeneurship. Thank you Clay Clark and Matthew Collin at the Genomic Sequencing Core for your expertise. Thank you to the administrative staff in graduate division for your assistance over the years.

While not included herein, I would like to acknowledge the following:

> Alva, T. R. et al. Efficient facemask decontamination via forced ozone convection. *Sci. Rep.* **11**, 1–11 (2021).

The co-first author Troy R. Alva designed and assisted the bacteria survivability study. The co-first author Joseph Schwan designed and manufactured the decontamination scheme. The co-authors Dr. Giorgio Nava and Dr. Carla Berrospe

Rodriguez assisted with the UV and FTIR characterization of the ozone concentration. The co-authors Dr. Pin Wang and Zachary Spencer Dunn assisted with the viral survivability study. The co-authors Dr. Joshua T. Morgan and Dr. Justin W. Chartron assisted with the interpretation of the survivability results. The co-authors Dr. Lorenzo Mangolini, Joseph Schwan, and Giorgio Nava conceived the decontamination scheme. The co-authors Troy R. Alva, Joseph Schwan, and Dr. Lorenzo Mangolini wrote the manuscript. We acknowledge Brian Lupish for technical assistance and Mario Leon Lopez for kindly providing the GFP expression plasmid.

The work described in this dissertation was funded by Bolt Threads Inc. (Emeryville, CA), the National Science Foundation (NSF CBET-1951942), and the Bourns College of Engineering at UC Riverside.

The text of this dissertation, in part, is a reprint of the material as it appears in:

> 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)

The co-authors Troy R. Alva and Dr. Justin W. Chartron designed experiments, performed analysis, and wrote the manuscript. The co-authors Troy R. Alva, Melanie Riera, and Dr. Justin W. Chartron performed experiments. The co-author Justin W. Chartron listed in that publication directed and supervised the research which partially forms the basis for this dissertation. We acknowledge Josh Kittleson (Bolt Threads), Gustavo Pesce (Abalone Bio), Thomas Stevens (Google), and Chris Love (MIT) for useful discussions as well as our colleagues in the Department of Bioengineering at UC Riverside.

To God, my family, and my closest friends.

Mom, this would not be made possible without the spark for learning that you instilled in me as a young boy. Thank you for your ever continuing encouragement.

Hannah, for your love, patience, and companionship. The fickleness of science, seemingly Sisyphean, would have been hard and arduous without the balance you gave to my life.

ABSTRACT OF THE DISSERTATION

Rational Designs for Increased Bioproduction in *Komagataella phaffii*

by

Troy Alva

Doctor of Philosophy, Graduate Program in Bioengineering
University of California, Riverside, September 2021
Dr. Ian Wheeldon, Chairperson

The production of biologics such as enzymes, biomaterials, and therapeutics play a leading role in biotechnology. As a microbial cell factory, *Komagataella phaffii* stands out for its high secretion capacity, ability to metabolize methanol as its primary carbon source, safety record as a source of biologics, and extensive literature compared to other non-model yeasts. Large scale production of biologics is simplified if the host secretes the proteins into its medium. However, this is difficult to achieve because an organism's protein biogenesis machinery have evolved under the demands of its proteome and cells struggle to express and secrete non-native proteins. Heterologous protein production requires the harmonization of biogenetic machinery like ribosomes for synthesis, signal recognition particles and their cognate receptor for intracellular targeting, and protein folding chaperones for post-translational modification. These biogenetic components represent limited pools of resources that are distributed between heterologous proteins and the host proteome. While in use, these components are sequestered and unavailable for other tasks. Accurate accounting of these resources allows strains to be engineered in ways that relieve bottlenecks specific to producing particular heterologous proteins.

We survey the translational landscape of *Komagataella phaffii* using Ribo-seq

under different conditions to elucidate which host proteins sequester the most biogenetic resources. Ribo-seq is a high throughput sequencing technique used to monitor protein synthesis by measuring ribosome abundancies at each codon of each transcript. Herein, a novel Ribo-seq pipeline was used to prepare mRNA footprint libraries for Illumina sequencing, create new annotations for protein-encoding genes, and address biases that are inherent to the technique and complicate quantification of protein synthesis. Using this pipeline, Ribo-seq was used with subcellular fractionation techniques to measure translation in the cytosol and on the endoplasmic reticulum membrane to uncover how and what proteins traffic through the early secretory pathway of yeasts. Finally, this developed protocol was used to identify which host proteins sequester the most biogenetic resources during heterologous expression. Genes encoding these proteins represent targets for rational strain engineering.

Table of Contents

List of Figures

List of Tables

Introduction

The market for engineered proteins has concomitantly increased with the expansion of protein production platforms.[1] Recombinant proteins are enormously useful and diverse technologies that range from enzymes, structural proteins including those found in spider silk, and proteins useful for therapeutic design. The demands for these technologies has led to the development of specifically tailored cellular chassis toward their production. The choice of one cell type over another requires the consideration of many factors including protein structural complexity, codon biases, glycosylation patterns, and minimum titer requirements. Cellular factories used for bioproduction are *E. coli*, yeasts, or mammalian cells.[2]

As microbial cell factories, yeasts offer many advantages for recombinant protein production including their natural properties and potential in synthetic biology. Yeasts grow rapidly to high densities in inexpensive media and are resistant to physical and chemical stress.[2] They also have an endomembrane system that is fundamentally conserved with higher eukaryotes.[3] This oxidative environment supports glycosylation and subsequent glycan modification, folding using ATP-driven molecular chaperones and protein disulfide isomerases, and protein quality control.[4] Compared to mammalian cells, yeasts have simpler genomes and can be more easily characterized and modified.[5] Combine these features with tools such as CRISPR/cas9, and the range of tractable species is expanding.[6,7]

Komagataella phaffii (one of two species previously known as *Pichia pastoris*)[8–10] stands out as a host for recombinant protein expression due to its high secretion capacity, its ability to metabolize methanol as its primary carbon source, its safety record as a source of biologics, and its extensive literature compared to other non-model yeasts.[11,12] As a methylotroph, *K. phaffii* contains a promoter region (*AOX1*) that constitutively regulates the expression of alcohol oxidase for methanol metabolization. This methanol induced promoter is used to control the production and secretion of heterologous proteins.[13] Thus, *K. phaffii* is an ideal chassis to rapidly implement changes designed to improve protein expression and secretion.[5] Indeed, recent work in *K. phaffii* has focused on systems-level analysis[14] and implementing design approaches of synthetic biology such as molecular parts lists and strain engineering.[15,16] Such changes may accelerate product development and allow cheap, local production of pharmaceuticals.[17,18]

Production of heterologous proteins is difficult but greatly simplified when heterologous proteins are secreted into the growth media.[2] For secreted proteins using the *AOX1* promoter, the first bottleneck in secretion appears to be translocation from the cytoplasm into the endoplasmic reticulum (ER) lumen.[19,20] This process is complex and requires the orchestration of many biogenesis factors. These factors mediate targeting to the membrane of the ER, translocation across this membrane, protein folding, post-translational modification, quality control, and trafficking. Membrane targeting during secretion largely occurs co-translationally and is contingent on the recognition and binding of N-terminus hydrophobic motifs, signal sequences, by a signal recognition particle (SRP).[21] SRP guides the ribosome nascent chain complex to the membrane of the ER where they associate with *sec* translocon complexes by interaction of SRP's cognate receptor.[22]

In yeasts, several varieties of *sec* translocons exist that are distinguished by their

subunit composition. Yeasts have two homologs of the central pore complex, Sec61 and Ssh1 complexes. The Sec61 complex is composed of Sec61p, Sss1p, Sbh1p and the Ssh1 complex is composed of Ssh1p, Sbh2p, and Sss1p. The Sec61 complex can also associate with Sec62p, Sec63p, Sec71p and Sec72p, while the Ssh1 complex cannot. While *Sec71* and *Sec72* are non essential genes, *in vivo* tagging of ribosomes reveals that deleting Sec71p reduces co-translational attachment of a subset of mRNAs.[23] Sec63p and Sec62p are essential for cell survival and have different roles depending on the translocation pathway. Both Sec61 and Ssh1 complexes can associate with the SRP receptor for co-translational translocation with Sbh1p and Sbh2p respectively. SRP dependent and independent pathways use different *sec* translocons for translocation and use different mechanisms for associating the ribosomal nascent chain complex (RNC) with the ER membrane.

During SRP dependent translocation, ribosomes exposing the signal peptides of secreted proteins are associated with and guided by SRP to the SRP's cognate receptor on the ER where the nascent protein translocates coincidentally with translation.[22] Co-translational translocation in *S. cerevisiae* occurs equally between the Sec61 hexameric translocon (Sec61 complex, Sec63p, Sec71p, and Sec72p) and the Ssh1 heterotrimeric translocon (Ssh1 complex). While a preponderance of evidence suggests that mRNA is physically attached to the ER, it remains unclear which states have ribosomes in contact with translocons *in vivo*. Post-translational translocation is an SRP independent process where signal sequences are directly recognized and associated by the heptameric translocon complex. Post-translational translocation uncontroversially occurs with tail-anchored membrane proteins and small secreted proteins. The post-translational heptameric translocon is composed of the Sec61 complex, Sec62p, Sec63p, Sec71p, and Sec72p. Sec63p plays conserved and unique roles for the hexameric (co-translational) and heptameric (post-translational) com-

plexes. Sec63p is speculated to interact with the ER resident chaperone Kar2p (BiP) for translocon pore gating. In the hexameric complex, Sec63p is necessary for forming the hexameric complex and is shown to be necessary for membrane proteins, which are SRP-dependent co-translational.[24] In the heptameric complex, Sec63p has been shown to block ribosomes from binding to Sec61p[25] and is necessary for complex assembly as it binds to Sec62p. Sec62p is a critical subunit in the heptameric complex as it directly recognizes and binds signal peptides for SRP-independent translocation.

Translocation depends on hydrophobic motifs: signal sequences, signal anchors or internal transmembrane domains. These motifs are recognized by either a translocon, or by the signal recognition particle (SRP).[21] Prior investigation revealed extreme cases where a protein strictly requires a specific translocon architecture or subunit for translocation. One extreme includes proteins which require the signal recognition particle for translocation, and are therefore obligate co-translational substrates. In yeast, obligate co-translational substrates are often vacuolar proteins, extracellular/secreted proteins, or membrane proteins. However, there are many examples of proteins which seem to rely on more than one translocon pathway. Proteins necessary for protein biogenesis such as Kar2p, Och1p, and Ost1p are transported by either pathway and are speculated to have evolved this way due to the essentiality of their functions. The preference for either of these pathways are consequence of the hydrophobicity of the signal sequence and is not easily predicted. This uncertainty in linking protein sequence to translocon requirement may hinder attempts to rationally engineer expression systems.

Translocation is facilitated with the help of molecular chaperones that are also involved in oxidative protein folding. Proteins that do not translocate sequester cytosolic chaperones linked to translocation, like *Ssa1*, *Ssa2*, *Ssb1* and *Ssb2*. While these

chaperones have been demonstrated to assist in translocation in normal conditions, we predict that overexpression during bioproduction will generate an artificially destabilized state. Ultimately, secreted proteins that fail to translocate into the ER do not have access to ER-resident chaperones and fail to fold correctly. Protein folding is an ATP driven process that includes many ER-resident proteins to *Kar2* such as *Scj1*, *Pdi1*, *Ero1*, and *Jem1*. Proteins that fail to assemble activate the unfolded protein response (UPR) and are retroactively translocated and degraded by the ER-associated degradation pathway (ERAD).[3] Highly expressed host proteins that are co-translationally translocated are expected to block translocation of heterologous proteins as Sec-translocons become limited in number and processivity. On the other hand, highly expressed proteins that are post-translationally translocated inhibit protein folding for heterologous proteins as protein folding chaperones are also limited in number and processivity. These genes that pose the greatest threat to ER trafficking are not well understood and may provide novel targets for maximizing heterologous flux in relation to the host proteome.

The Lewis lab has elegantly demonstrated that deleting highly expressed concomitant heterologous proteins in CHO cells increases the yield of other heterologous secreted proteins.[26] This result is intuitive, as highly expressed proteins compete for resources necessary for protein secretion. Other studies have also shown positive, yet variable, results when modifying the host's secretory system. Modification of the secretory pathway, such as the optimization of signal sequences for protein targeting[27] and reducing the effect of the ERAD system,[20] provides varying degrees of success and is contingent on the complexity and specialization of the protein product.[28,29] We hypothesize that the discrepancy in efficacy between viable secretory modifications is because a cell's biogenesis machinery has coevolved under the demand of its proteome.

Choosing gene targets to improve secretion in non-model yeast is not trivial. While fungi are genetically and physiologically diverse, most genetic manipulations are derived from knowledge of baker's yeast, the model organism *S. cerevisiae*.[3] This is problematic when strain engineering *K. phaffii*, however, as approximately 230 million years of evolutionary divergence separates these two species. The large range of genetic diversity between species results in subtle differences in both protein sequence and regulation of gene expression, which can cause drastically different phenotypes in regards to protein production and secretion.[3] In addition to differential phenotypes in protein synthesis between species, the majority of next generation sequencing data for *K. phaffii* is derived from cells cultured in media distinct from those used in heterologous conditions. Currently, it is unclear how the translatome of *K. phaffii* differs between these conditions and if gene targets for strain engineering consequently change.

Thus, rational engineering of the host's secretome requires a product-tailored approach that considers inherent metabolic burdens as well as the demands that protein sequence features require for proper targeting and secretion. Proteins that are longer will require more ribosomes to produce the same amount of proteins than a shorter gene may require in a given amount of time and may benefit from deleting ribosome rich transcripts from the host.[30] Longer proteins also have complicated folding requirements and may require a greater number of protein folding chaperones than a shorter protein would.[31] Shorter proteins may have adequate ribosomal resources but may compete for translocon accessibility in the ER. Understanding which genes require and sequester the most biogenetic resources for production requires a set of techniques that provides genome wide coverage and accurately reflects demands on biogenetic resources.

Ribosome profiling (Ribo-seq) is a high throughput sequencing technique used

to monitor protein synthesis by measuring ribosome abundancies at each codon of each transcript. This technique has advantages over standard proteomics in its ability to detect expression with wider gene coverage and is more accurate in detecting protein abundance in comparison to RNA-seq.[32] Ribo-seq is a variation of RNA sequencing, where a non-specific ribonuclease generates varying sized "footprints" of ribosome-protected mRNA depending on the translational state of the ribosome.[33] In these regards, Ribo-seq can provide an instantaneous snapshot into the translatome and represents an accurate bridge between gene expression and protein abundance.[34] Combined with fractionation techniques to separate cytosolic and ER membrane-bound ribosomes, this procedure allows us, for the first time, to connect sequence features to membrane expression as well as survey metabolic burdens that impede protein production.

The first chapter of this dissertation presents the development of a Ribo-seq protocol for the characterization of *K. phaffii*'s translatome. Herein, a Rib-Seq protocol was used to prepare mRNA footprint libraries for Illumina sequencing, create new annotations for protein-encoding genes, and address biases inherent to Ribo-seq in this species. Interpreting Ribo-seq studies in *K. phaffii* is made difficult as previous genome annotations do not have accurate demarcations of untranslated regions (UTRs) and open reading frames (ORFs). We utilize Ribo-seq data and long-read RNA sequencing data to generate a novel annotation of protein-encoding genes in *K. phaffii*. While these annotations greatly improved our understanding of the translatome, assigning reads remains difficult as *K. phaffii*'s genome is tightly packed and small footprints have the capacity to map to multiple locations in the genome. We introduce a technique to ameliorate multi-mapping with a unique masking strategy that prevents codons with the highest propensity to map to homologous regions. This technique utilizes a novel metagene normalization procedure to more

accurately count reads that lie within homologous regions. Finally, we use this technique to discover which genes are most highly expressed in *K. phaffii* in YPD culture.

The second chapter of this dissertation uncovers which proteins require the most biosynthetic resources in the early secretory pathway of yeasts. We use our Ribo-seq protocol discussed in chapter one as well as subcellular fractionation to measure protein synthesis in the cytosol and on the surface of the ER-membrane. This analysis allows us to classify proteins that enter the secretory pathway co-translationally and predict those that enter post-translationally. Our libraries reveal that a subset of secreted proteins show less partitioning of their encoding to the surface of the ER. We show that nearly all of these proteins have relatively fewer amino acids between their targeting signals and C-termini and that kinetics of translation influence partitioning of mRNA to the ER. For co- and post-translational pathways, we estimate each protein's demand for ribosomes, translocons and molecular chaperones with genome wide coverage. For the first time, we compare the early secretory demands between different species. Using *K. phaffii* and *S. cerevisiae*, we show that a distinct set of proteins enter the ER and a strain specific understanding of the secretory pathway is necessary to rationally engineer cells for increased bioproduction.

The third chapter of this dissertation investigates *K. phaffii*'s translational landscape under conditions used for heterologous expression. In this chapter, we use Ribo-seq to model protein synthesis ER flux before and after heterologous induction. For heterologous expression, we utilize common techniques in industry where cells are grown in glycerol media before induced to express heterologous proteins using methanol media. This study is the first of its kind to capture differences in host protein synthesis and early secretory trafficking consequent to over-expressing and

secreting recombinant proteins. Herein, we further optimize Ribo-seq for surveying translation of cells in different conditions and draw several conclusions to rationally engineer cells for increased bioproduction. In optimizing Ribo-seq, we show that ribosomal rRNA depletion strategies are more efficient when they are designed under similar conditions to those used for Ribo-seq. We also develop a novel technique using biological replicates to determine minimum read count thresholds when comparing expression between different samples. To rationally engineering cells for increased bioproduction, we identify which host proteins sequester the most biogenetic resources and use these to provide rational targets for designing strains for increased biogenesis. These experiments highlight that gene targets used for strain engineering change for cells grown in different conditions and that cells' proteome responds to heterologous expression in unpredictable ways.

Chapter 1

Optimization of ribosome profiling in *Komagataella phaffii*

Background: A segment of industrial bioproduction relies on *Komagataella phaffii* as a production host for its growth characteristics, ability to produce proteins with complex folding and glycosylation requirements, and ability to utilize methanol as a sole carbon source. While its use in industry continues to grow, however, little is known of the organism's translational landscape. Furthermore, efforts to understand *Komagataella phaffii*'s translatome are complicated by idiosyncrasies related to its genome and the tools used for its interrogation. Herein, we make general and strain specific optimizations to high throughput sequencing techniques aimed to answer fundamental questions related to protein synthesis.

Results: In optimizing ribosome profiling, we adapt library preparation techniques to study translation in *Komagataella phaffii* during exponential growth in rich media with unprecedented resolution and throughputness. By coupling ribosome profiling with long-read mRNA sequencing, we generated a new annotation of protein-encoding genes. Next, we developed masking strategies to ameliorate the

propensity for ribosome footprints to map to multiple locations in *Komagataella phaffii*'s tightly packed genome. Masking strategies involved read distribution normalizations that correct for mask position biases secondary to common ribosome profiling library preparation artifacts. These strategies allowed us to calculate protein synthesis and ribosome sequestration metrics per gene.

Conclusion: Ribosome profiling is an incredibly useful tool to understand protein synthesis dynamics. Our pipeline offers a systematic approach to study translation and improve existent genome annotations.

1.1 Introduction

To provide a model for the translational landscape of *Komagataella phaffii*, we used ribosome profiling (Ribo-seq). Ribo-seq is a high throughput sequencing technique used to measure ribosome abundance at each codon of each transcript. Compared to standard RNA-seq experiments, Ribo-seq more accurately detects static protein levels.[32] While actively translated mRNA read counts do not reflect proteostasis levels, we are interested in quantifying protein synthetic demands and so Ribo-seq is a more appropriate tool than standard proteomics. Indeed, this technique has additional advantages over standard proteomics for its ability to detect expression with wider gene coverage[32] while accurately predicting protein stoichiometry.[35] In these regards, Ribo-seq represents an appropriate tool for bridging gene expression, protein synthesis demands, and protein abundance.[34]

Ribo-seq involves the capture of actively translating ribosome protected mRNA fragments in the cytosol and on the surface of membranes like the ER and mitochondria. During Ribo-seq, cell cultures are typically grown to log phase before they are collected and flash frozen.[36–38] Flash frozen cells are then lysed before non-specific

nuclease is used to degrade mRNA strands that are not covered, and thus protected, by ribosomes. While it is not difficult to separate the protein components of the ribosome after nuclease digestion, approximately $65\,\%$ of the ribosome is composed of RNA.[39] Ribosomal RNA (rRNA), however, requires measures greater than organic extraction procedures to separate from ribosome protected mRNA fragments. rRNA contaminates Ribo-seq libraries by obscuring the percentage of next generation sequencing reads that map to open reading frames (ORFs).[40] There are many ways to reduce rRNA contamination and the most common method is probe-directed degradation using biotinylated oligos.[36] Commercial kits that utilize this method are often limited and are prohibitively restricted to a subset of model organisms like *E. coli*, *S. cerevisiae*, mice, and Chinese hamster ovary cells.[41] While there is a degree of homology between *K. phaffii* and some of these organisms (notably *S. cerevisiae*), it is unclear if these kits are effective at removing rRNA contaminants for other organisms as their probes' sequences are often proprietary.[42]

Interpreting protein synthesis in *K. phaffii* from Ribo-seq derived data sets is made difficult by idiosyncrasies inherent to the technique and the species. Ribosome protected mRNA fragments are small and have the capacity to map to multiple locations in the genome.[35,43] Ribo-seq also has a tendency to disproportionately collect ribosomes localized at the the 5' end of transcripts due to slower rates of initiation than elongation, pausing, or as a byproduct of library preparation.[44,45] Together, these artifacts complicate protein synthesis calculations. Differential expression studies are further complicated for *K. phaffii* as genome annotations for the non-model organism's genes are not often correct. While *K. phaffii*'s genome and transcriptome have thoroughly been sequenced,[46–48] the boundaries of its ORFs are guided by *de novo* predictors and lack *in vivo* based evidence of actively translated regions. These *de novo* predictors may misannotate translational start sites and not

recognize very short ORFs.[49] While canonical methods do not exist to address these difficulties, super-resolution Ribo-seq profiles have been employed to annotate transcript ORFs with great success.[50,51]

The protocol described in this chapter optimizes Ribo-seq preparation and analysis in *K. phaffii*. In optimizing library preparation protocols, we adapt the methods of Jonathan Weissman[36] and Nicholas Ingolia.[37] The methodologies outlined herein are necessary to begin species and condition specific comparisons that can be used for rational strain engineering.

1.2 Materials

1.2.1 Strains

The dissertation herein uses the *Komagataella phaffii* strains shown in *Table 1.1*.

Table 1.1: **Strains used in Ribo-seq**

Strain	Genotype	Phenotype	Manufacturer	Catalog.No.
GS115	$his4$	Mut^{+}	Invitrogen	K1710-01
GS115 Albumin	$HIS4$	Mut^{S}	Invitrogen	K1710-01

1.2.2 Reagents

Ribo-seq requires a sizable variety of reagents. While many of these reagents may be replaced with equivalents, we have shown great success in the accuracy and replicability of our library preparation protocol using the ones listed in *Table 1.2*.

Table 1.2: **Reagents used in Ribo-seq**

Material	Manufacturer	Catalog.No.
RNase inhibitors		
RNase-free water	Invitrogen	AM9930
SUPERase*In 20 U/ul	Invitrogen	AM2694
Liquid nitrogen[a]	AirGas	13600-102
Ribo-seq chemicals		
100 mg/ml Cycloheximide[b]	Sigma–Aldrich	C4859-1ML
Ambion RNase A 1 mg/ml	Invitrogen	AM2270
Buffers and sucrose gradients		
NaCl, RNase-free	Fisher Scientific	S271-1
0.5 M EDTA, RNase-free	Gentrox	30-012
10% SDS, molecular grade	Promega	V6551
NaOAc, RNase-free	Gentrox	30-038
MgOAc, RNase-free	RPI	M24100-500.0
MOPS, RNase-free	RPI	M92020-500.0
KOH, RNase-free	Fisher Scientific	SP208-500
KOAc, RNase-free	Fisher Scientific	BP364-500
Triton X-100, molecular grade	RPI	111036
Tris-HCl, molecular grade	Fisher Scientific	PR-H5123
Tris Base, molecular grade	Gentrox	30-065
Sucrose, molecular grade	Ricca Chemical	RSOC0020-1C
Precipitations		
Oligo Clean & Concentrator kit	Zymo Research	D4060
PCR Purification Kit, Agencourt AMPure XP	Fisher Scientific	BP220-1
Phenol:chloroform 5:1, molecular grade[c]	Sigma-Aldrich	P1944-100ML
Chloroform, molecular grade[c]	Beckman Coulter	A63880
Isopropanol, molecular grade[d,e]	Ricca Chemical	RSOI0020-1C
Ethanol, molecular grade[f]	Fisher Scientific	BP2818-4
GlycoBlue 15 mg/ml	Invitrogen	AM9515
Gel electrophoresis		
SureCast Acrylamide Solution (40%)[g]	Invitrogen	HC2040
10x TBE RNase-free	RPI	T32020-10000.0
Bromophenol blue	Fisher Scientific	BP115-25
Formamide, molecular grade[i]	Promega	H5051
10000x SYBR Gold[h]	Invitrogen	S11494
Ultra Low Range DNA Ladder, 10-300 bp	Invitrogen	10597012
VersaLadder, 100-10,000 bp	Goldbio	D012-500
Ribosomal RNA depletion		
Ribo-Zero Gold rRNA Removal Kit (Yeast)	Illumina	MRZY1306
Duplex-specific nuclease	Evrogen	EA001
Exonuclease I	New England Biolabs	M0293S
Dephosphorylation and linker ligation		
T4 polynucleotide kinase	New England Biolabs	M0201L
T4 RNA Ligase 2 truncated K227Q	New England Biolabs	M0351L
Mth RNA Ligase	New England Biolabs	M2611A
Reverse transcription		

Table 1.2: **Reagents used in Ribo-seq** *(continued)*

Material	Manufacturer	Catalog.No.
10 mM dNTP mix	New England Biolabs	N0447S
SuperScript II	Invitrogen	18064014
NaOH[j]	RPI	S24000-500.0
Circularization and amplification		
CircLigase II	Lucigen	CL9025K
Q5 High-Fidelity DNA Polymerase	New England Biolabs	M0491L

[a] CAUTION - Liquid nitrogen may cause burns or suffocation
[b] CAUTION - Cycloheximide is extremely hazardous, dispose properly and handle with care
[c] CAUTION - Extremely hazardous, use in fume hood
[d] CAUTION - Highly flammable and volatile
[e] CAUTION - Irritant
[f] CAUTION - Acrylamide is a neurotoxin, handle with care
[g] CAUTION - Nucleid acid stains are mutagenic, dispose properly and handle with care
[h] CAUTION - Formamide is a reproductive toxin
[i] CAUTION - Highly corrosive

1.2.3 Oligonucleotides

The oligonucleotides shown in *Table 1.3* are as described in Ingolia et al..[37]

Table 1.3: **Oligonucleotides for library preparation**

Oligo Id	Oligo sequence
Linker sequences	
NI-810	5′-/5Phos/NNNNNATCGTAGATCGGAAGAGCACACGTCTGAA/3ddC/
NI-811	5′-/5Phos/NNNNNAGCTAAGATCGGAAGAGCACACGTCTGAA/3ddC/
NI-812	5′-/5Phos/NNNNNCGTAAAGATCGGAAGAGCACACGTCTGAA/3ddC/
NI-813	5′-/5Phos/NNNNNCTAGAAGATCGGAAGAGCACACGTCTGAA/3ddC/
NI-814	5′-/5Phos/NNNNNGATCAAGATCGGAAGAGCACACGTCTGAA/3ddC/
NI-815	5′-/5Phos/NNNNNGCATAAGATCGGAAGAGCACACGTCTGAA/3ddC/
Reverse transcription primer	
NI-802	5′-/5Phos/NNAGATCGGAAGAGCGTCGTGTAGGGAAAGAG/iSp18/GTGACTGGAGTTCAGACGTGTGCTC
Forward library PCR primer	

Table 1.3: **Oligonucleotides for library preparation** *(continued)*

Oligo Id	Oligo sequence
NI-NI-798	5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC
Reverse library PCR primers	
NI-799	5′-CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTG
NI-822	5′-CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTG

1.3 Methods

1.3.1 Culture conditions, harvesting, and lysis

Assays described were performed using *Komagataella phaffii* GS115 Mut$^+$ cultured in YPD ($1\,\%$ yeast extract, $2\,\%$ peptone and $2\,\%$ glucose).

1. For each Ribo-seq biological replicate, grow liquid culture to an OD_{600} of 2 at $30\,°C$ with shaking in baffled $2\,L$ flasks.
2. Vacuum filter cells from culture using a $0.8\,\mu m$ filter.
3. Immediately after filtering, scrape cells off the filter using a chilled scoopula before submerging in a $50\,mL$ conical tube containing liquid nitrogen. Samples may be indefinitely stored at $-80\,°C$.
4. Prepare lysis buffer containing $50\,mM$ MOPS, $25\,mM$ KOH, $100\,mM$ KOAc, $2\,mM$ MgOAc, $1\,mM$ DTT, $1\,\%$ Triton X-100, and $100\,\mu g\,mL^{-1}$ CHX when appropriate.
5. Freeze lysis buffer by adding $2\,mL$ per dropwise to a $50\,mL$ conical tube containing liquid nitrogen.

6. For each biological replicate, mix frozen cells with $2\,\mathrm{mL}$ frozen lysis buffer into single $50\,\mathrm{mL}$ ball mill chamber (Retsch) with a single $2\,\mathrm{mL}$ steel ball (Retsch).
7. Pulverize mixture for $2\,\mathrm{min}$ and collect into $50\,\mathrm{mL}$ conical tubes.
8. After thawing, centrifuge lysates at $18\,000\,\mathrm{g}$ for $10\,\mathrm{min}$.
9. Transfer supernatants to $1.5\,\mathrm{mL}$ conical tube before further clarification via centrifugation at $23\,000\,\mathrm{g}$ for $20\,\mathrm{min}$.

1.3.2 Ribo-seq

1.3.2.1 Nuclease footprinting and ribosome recovery

1. Determine the rough RNA concentration of cell lysates using BioDrop UV spectrometer. Dilute samples in lysis buffer to similar concentrations.
2. Prepare $10\,\%$ to $50\,\%$ sucrose gradients in $50\,\mathrm{mM}$ Tris pH 7.5, $200\,\mathrm{mM}$ NaCl, and $2\,\mathrm{mM}$ MgOAc using a Gradient Master (Biocomp). Store at $4\,^{\circ}\mathrm{C}$ for $1\,\mathrm{h}$.

3. Determine ideal RNase A concentrations by using different concentrations for multiple $300\,\mu\mathrm{L}$ lysates. For each lysate, digest with RNase A for $1\,\mathrm{h}$ at room temperature before layering digested samples on sucrose gradients. Centrifuge at $39\,000$ RPM for $2.5\,\mathrm{h}$ in a TH-641 rotor (Thermo). After centrifugation, fractionate gradients using a Piston Gradient Fractionator (Biocomp) to determine how many units of RNase A are necessary for optimizing monosome peaks (*Figure 1.1*).
4. Repeat digestions using determined units of RNase A described above, retain monosome peaks.
5. Extract RNA using a standard phenol-chloroform method followed by isopropranol precipitation

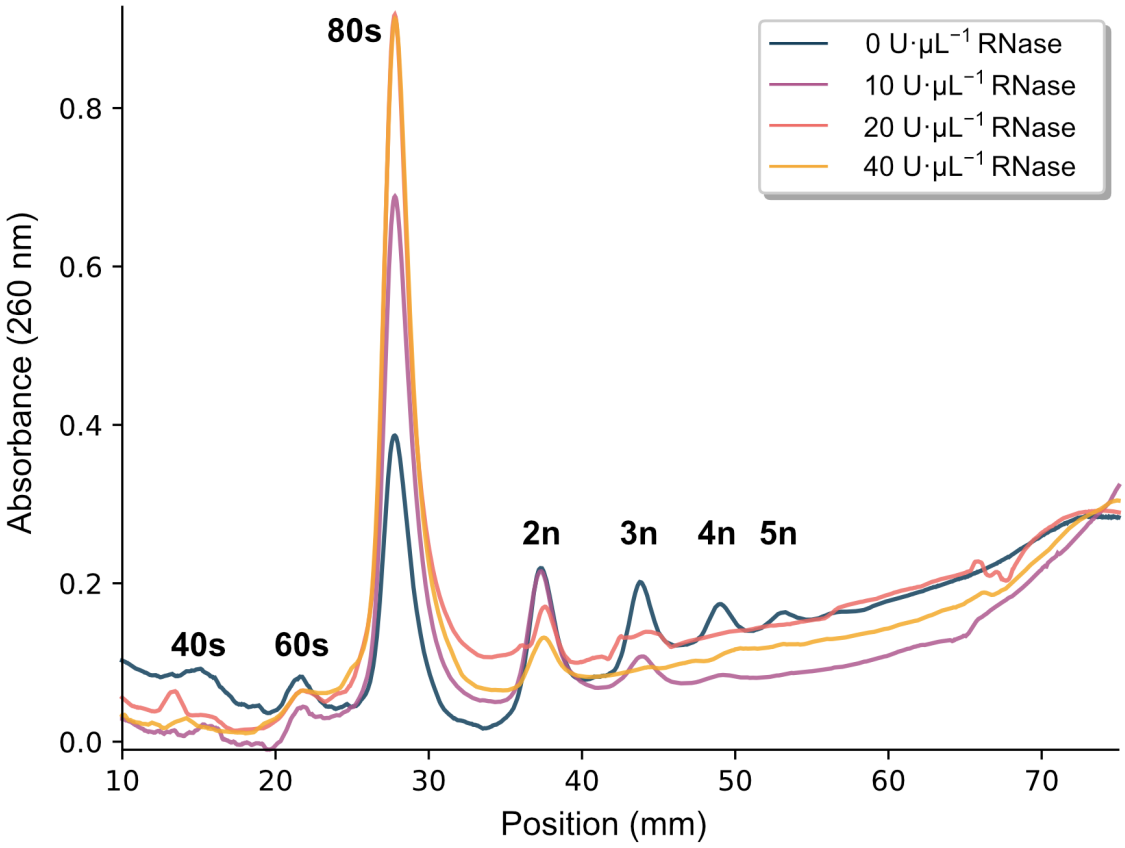

Figure 1.1: **Determination of ideal RNase concentration for nuclease footprinting** Images represent absorbance reading from Piston Gradient Fractionator (Biocomp). For *K. phaffii* lysates at this concentration, 40 U RNase A maximizes the area under the curve for 80s monosome peaks relative to the rest of the gradient.

6. Purify ribosome protected footprints $18\,\mathrm{nt}$ to $34\,\mathrm{nt}$ using $15\,\%$ polyacrylamide TBE-urea gel (*Figure 1.2*).
7. Perform over night extraction of excised gel fragments using RNA gel extraction buffer followed by isopropranol precipitation and resuspension in $4\,\mu\mathrm{L}$ water containing $20\,\mathrm{U\,mL^{-1}}$ SUPERase · In (1:1000 SUPERase · In).

1.3.2.2 Dephosphorylation and linker ligation

1. Dephosphorylated purified fragments by incubating $2\,\mu\mathrm{L}$ $1\,\mathrm{M}$ RNA sample with $2\,\mu\mathrm{L}$ RNase free water, $0.5\,\mu\mathrm{L}$ SUPERase · In RNase Inhibitor, $0.5\,\mu\mathrm{L}$ T4

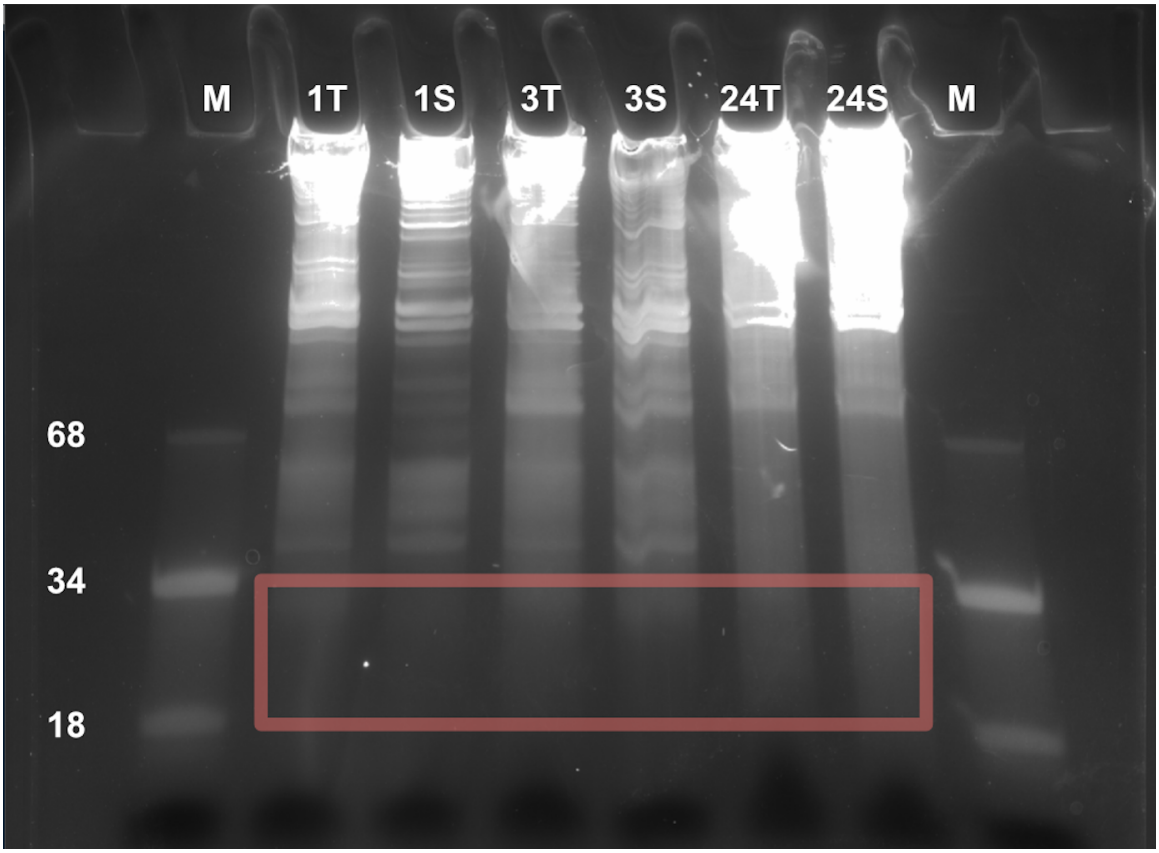

Figure 1.2: **Ribosome footprint size selection** The "M" on the outside lanes is an RNA marker composed of $18\,\mathrm{nt}$ and $34\,\mathrm{nt}$ RNA fragments. The lanes in the middle are samples. Footprint fragments are purified by excising the gel areas within orange box.

Polynucleotide Reaction Buffer (PNK), and $0.5\,\mu\mathrm{L}$ T4 Polynucleotide Kinase at $37\,^{\circ}\mathrm{C}$ for $1\,\mathrm{h}$.

2. Pre-adenylate adapter sequences by combining $1.2\,\mu\mathrm{L}$ $100\,\mu\mathrm{M}$ linker oligonucleotide with $2\,\mu\mathrm{L}$ 10x 5′ DNA adenylation reaction buffer, $2\,\mu\mathrm{L}$, $2\,\mu\mathrm{L}$ $1\,\mathrm{mM}$ ATP, $13.8\,\mu\mathrm{L}$ RNase-free water, and $2\,\mu\mathrm{L}$ M^{th} RNA ligase. Incubate for $1\,\mathrm{h}$ at $65\,^{\circ}\mathrm{C}$ followed by heat inactivation of enzyme via incubation at $85\,^{\circ}\mathrm{C}$ for $5\,\mathrm{min}$. Add $30\,\mu\mathrm{L}$ RNase-free water to adenylated adapter sequences followed by purification via Oligo Clean & Concentrator kit, elute in $6\,\mu\mathrm{L}$ 1:1000 SUPERase · In.
3. Combine $5.5\,\mu\mathrm{L}$ purified fragments, $3.5\,\mu\mathrm{L}$ $50\,\%$ PEG-8000, $0.5\,\mu\mathrm{L}$ 10x T4 RNA

Ligase Reaction Buffer, 0.5 µL 10 µM pre-adenylated adapter sequences and 0.5 µL T4 Rnl2(tr)k277Q. Incubate reaction mixture at 30 °C for 4 h.

4. Add 210 µL RNase-free water, 30 µL 3 M NaOAc, to each sample. Concentrate samples via isopropranol precipitation and resuspend in 3 µL 1:1000 SUPERase · In.
5. Purify linker-ligated samples using 15 % TBE-urea polyacrylamide gel.
6. Perform over night extraction of excised gel fragments using RNA gel extraction buffer followed by isopropranol precipitation.
7. Resuspend samples in 28 µL 1:1000 SUPERase · In. Using the images from previous gel purification, dilute and pool samples to equivalent concentration using their relative pixel intensities calculated from BioRad imaging software (*Figure 1.3*). Be sure to note which samples were linker ligated with which adapter sequences. Do not pool samples together that were linker ligated with the same adapter sequences as there is no way to computationally demultiplex them.

1.3.2.3 Depletion of ribosomal RNA using Ribo-Zero rRNA Removal Kit Yeast

1. Remove ribosomal RNA using streptavidin-coated magnetic beads from the Ribo-Zero rRNA Removal Kit Yeast as recommended by manufacturer.
2. Combine 90 µL RNase-free water and 18 µL 3 M NaOAc with samples before concentration via isopropranol precipitation.
3. Purify linker-ligated samples using 15 % TBE-urea polyacrylamide gel.
4. Perform over night extraction of excised gel fragments using RNA gel extraction buffer followed by isopropranol precipitation.
5. Resuspend samples in 10 µL 1:1000 SUPERase · In.

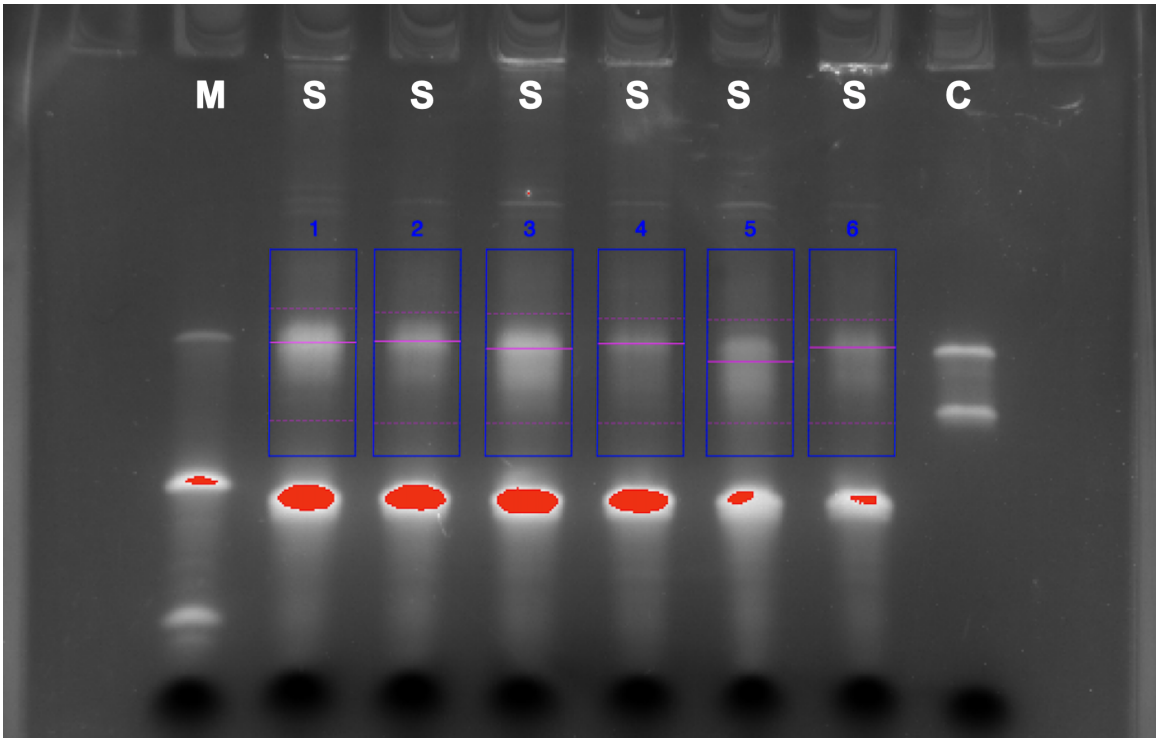

Figure 1.3: **Pooling linker-ligated RNA fragments** The "M" on the left-most lane is an RNA marker composed of 18 nt and 34 nt RNA fragments. The "S" lanes in the middle are samples. Sample boxes are generated by BioRad imaging software to determine relative concentration by pixel intensity. The "C" on the right-most lane is a linker-ligated RNA marker.

1.3.2.4 Reverse transcription

1. Perform 5 min denaturation of 10 µL sample via incubation at 65 °C with 2 µL reverse transcription primer NI-802.
2. Reverse transcribe samples by combining 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 before 30 min incubation at 50 °C.
3. Deactivate enzyme via hydrolysis by adding 2.2 µL 1 M NaOH followed by a 20 min incubation at 70 °C.
4. Add 28 µL RNAse free water was added to reverse transcription mixture (~50 µL total) and concentrate using Oligoclean and Concentrator Kit (it is no

longer necessary to resuspend samples in 1:1000 SUPERase · In).

5. Purify cDNA from unincorporated reverse transcription primers using $12\,\%$ TBE-urea polyacrylamide gel (*Figure 1.4*).

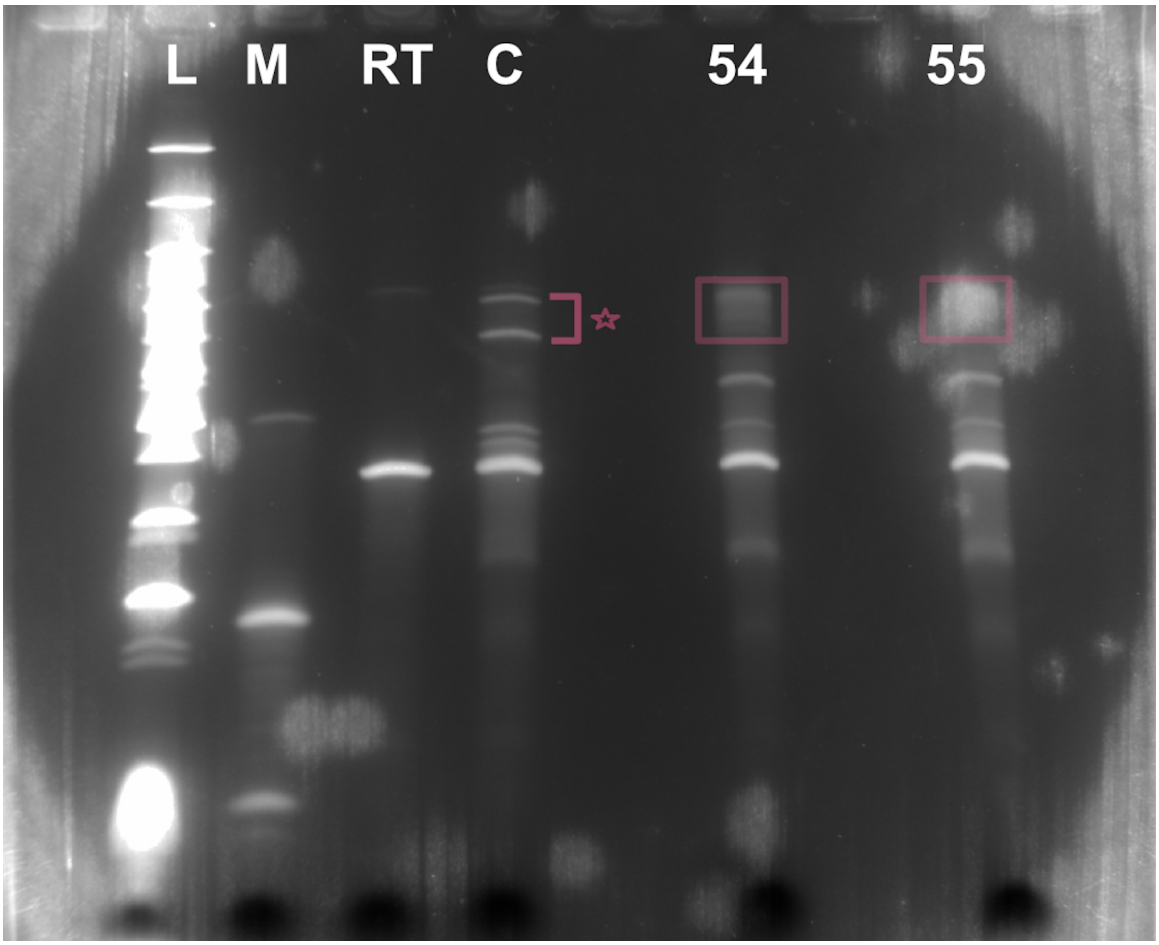

Figure 1.4: **Purifying reverse transcribed products** The "L" lane corresponds to an Ultra Low Range DNA ladder. The "M" lane corresponds to an RNA marker composed of $18\,\mathrm{nt}$ and $34\,\mathrm{nt}$ RNA fragments. The "RT" lane corresponds to the reverse transcription primer. The "C" lane corresponds to reverse transcribed RNA marker (previously linker ligated). The "54" and "55" lanes are samples.

6. Perform over night extraction of excised gel fragments using RNA gel extraction buffer followed by isopropranol precipitation.
7. Resuspend samples in $11\,\mu\mathrm{L}$ RNase-free water.

1.3.2.5 Circularization

1. Circularize single stranded cDNA samples by incubating $11\,\mu L$ sample in $2\,\mu L$ CircLigase II 10x Reaction Buffer, $1\,\mu L$ $50\,mM$ $MnCl_2$, $1\,\mu L$ ATP, $4\,\mu L$ $5\,M$ Betaine, and $1\,\mu L$ $100\,U\,\mu L^{-1}$ CircLigase II ssDNA Ligase at $60\,°C$ for $3\,h$.
2. Heat inactivate circularization process by incubating sample at $80\,°C$ for $10\,min$.

1.3.2.6 Probe-directed depletion of ribosomal RNA using DSN

The following process was utilized in Chapter 3 as the Ribo-Zero rRNA Removal Kit Yeast was insufficient for adequate rRNA depletion under those conditions. For cultures grown in YPD, this step may be skipped. Depletion probes were designed using rRNA aligned Ribo-seq reads collected from GS115 cultured in BMGY before methanol induction (*Table 3.1*).

1. Incubate $10\,\mu L$ circularized sample with $4\,\mu L$ 4x hybridization buffer, $1\,\mu L$ 4x depletion probes at $200\,\mu M$, and $1\,\mu L$ water.
2. Denature mixture for $2\,min$ at $98\,°C$ and allow to slowly anneal for $5\,h$ at $65\,°C$.
3. Enzymatically degrade rRNA fragments hybridized to depletion probes by adding $2\,\mu L$ 10x DSN master buffer, $1\,\mu L$ DSN storage buffer, and $1\,\mu L$ DSN before incubation at $65\,°C$ for $25\,min$.
4. Deactivate reaction by adding $20\,\mu L$ $10\,mM$ EDTA to mixture.
5. Purify samples using AMPure XP beads per manufacturer's instructions.
6. Digest linearized DSN degraded DNA fragments using Exonuclease I as these may contain regions complementary to PCR amplification primers.
7. Purify samples again using AMPure XP beads per manufacturer's instructions.

1.3.2.7 Library construction PCR

1. Determine the ideal number of rounds of PCR to reach appropriate concentrations without introducing amplification biases. Create reaction $50\,\mu L$ reaction mixture consisting of $10\,\mu L$ Q5 Reaction Buffer, $1\,\mu L$ $10\,mM$ dNTPs, $2.5\,\mu L$ $10\,\mu M$ forward primer, $4\,\mu L$ circularized DNA sample, $0.5\,\mu L$ Q5 High Fidelity DNA Polymerase and $29.5\,\mu L$ RNAse free water. Divide reaction mixture into 5x $10\,\mu L$ aliquots and perform different cycles of PCR for each aliquot (*Figure 1.5*).

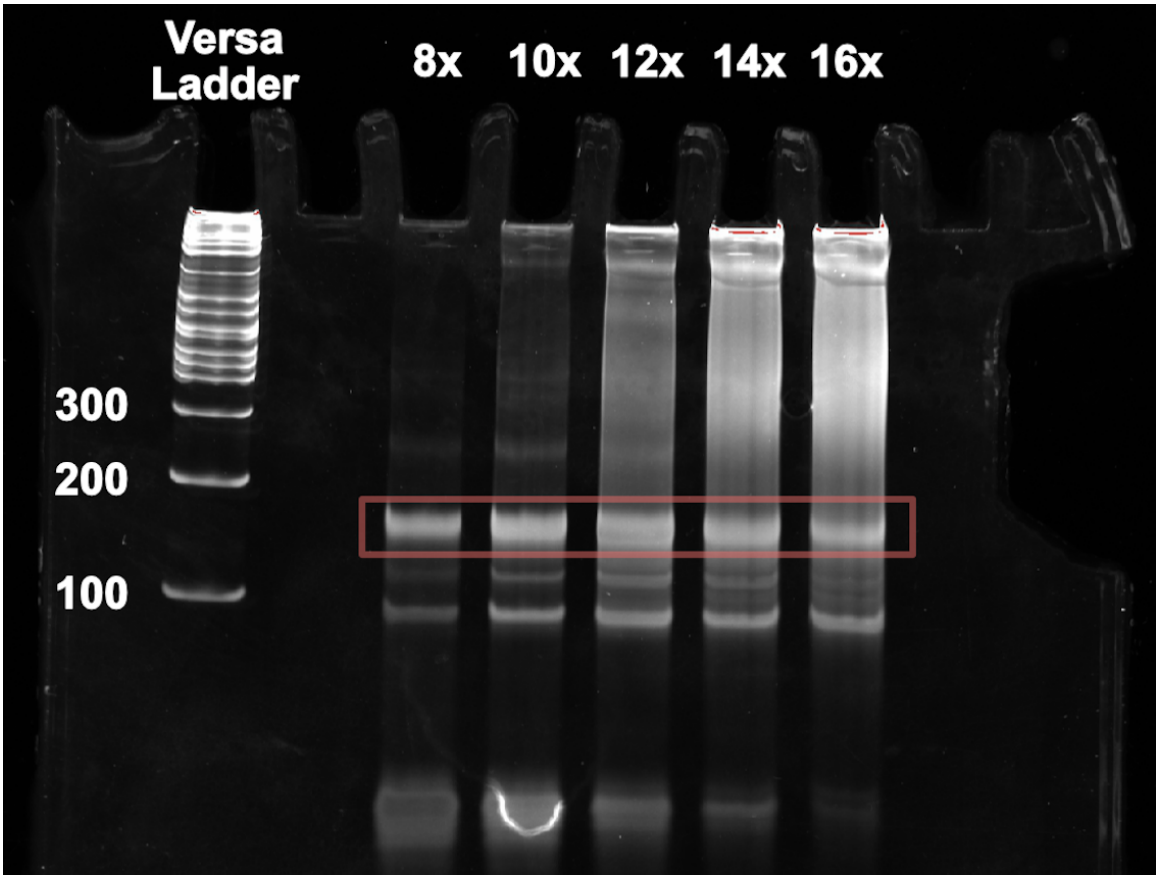

Figure 1.5: **Determining optimal conditions for library construction PCR** The left-most lane is a VersaLadder DNA ladder. The lanes labeled "nx" represent different amounts of PCR amplifications on technical replicates. For this sample, 10x rounds of amplification were required to maximize sample concentration with minimal re-annealed duplexes.

2. Amplify circularized DNA samples (split into five aliquots, again) using optimal conditions.
3. Purify DNA libraries from using $10\,\%$ non-denaturing TBE polyacrylamide gel. 4.Perform over night extraction of excised gel fragments using DNA gel extraction buffer followed by isopropranol precipitation.
4. Quantify library using Qubit 2.0 Fluorometer and dilute to required concentration.
5. Submit library for Illumina sequencing using HiSeq 4000 (Chapter 1 and Chapter 2) or NextSeq (Chapter 3).

1.3.3 Long read RNA sequencing

For each Ribo-seq biological replicate, $500\,\mathrm{mL}$ liquid cultures of YPD ($1\,\%$ yeast extract, $2\,\%$ peptone and $2\,\%$ glucose) were grown to an $OD_{600\,nm}$ of 2 at $30\,^{\circ}\mathrm{C}$ with shaking in baffled $2\,\mathrm{L}$ flasks and harvested via centrifugation. Total RNA was obtained using a Direct-Zol kit (Zymo Research). Cells were vortexed with glass beads for $2\,\mathrm{min}$ during incubation with TRI reagent. Total RNA was purified using Zymo Direct-zol columns and reagents according to the manufacturer's protocol. RNA was reverse transcribed using strand switching primer before selecting for full-length reverse transcribed RNA using PCR. Reverse transcribed RNA is amplified using PCR followed by a clean up using AMPure XP beads. Purified cDNA is then prepared for cDNA-PCR sequencing using Oxford Nanopore Technologies (ONT) minION sequencer[52,53] by adding an adapter to the amplified library. Flow cell for minION sequencer was then primed and loaded for long read RNA sequencing. Sequencing was performed using ONT's miniKNOW software.[54,55]

1.3.4 Transcript assembly

Novel transcripts were assembled using data derived from Ribo-seq, long-read RNA-Seq, and a prior genome sequence of strain GS115.[46] A flowchart of the annotation pipeline is provided in *Figure 1.6*. Ribo-seq reads and long reads were

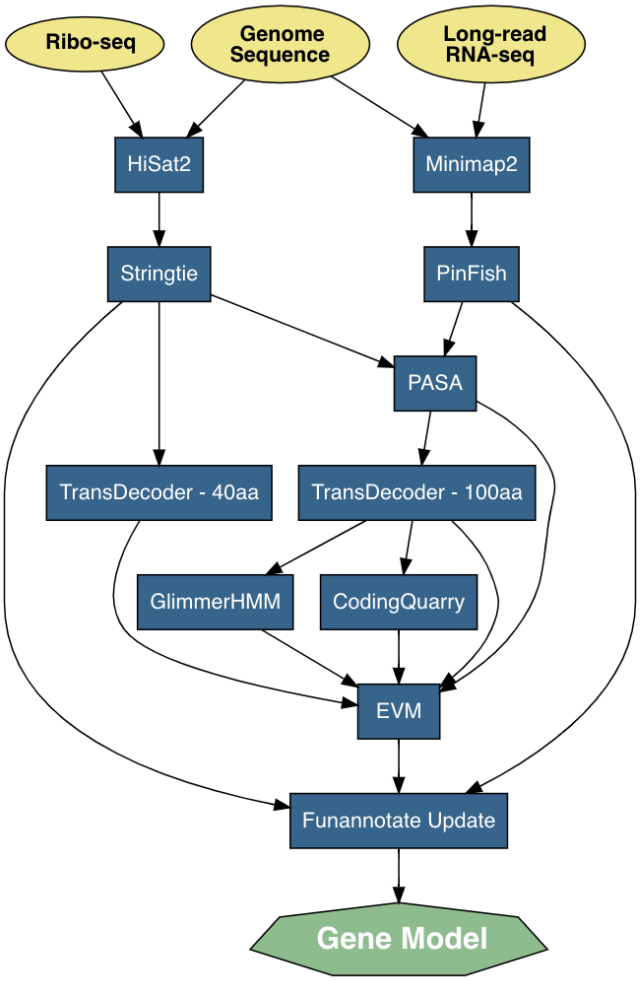

Figure 1.6: **Flow-chart of the annotation pipeline** Annotation pipeline involves Illumina sequenced ribosome profiling reads, Oxford Nanopore sequenced long-read RNA sequencing derived reads, and prior genome sequence of strain GS115.[46] Reads are processed and prepared for annotation preparation using several programs.

aligned to the reference genome using HISAT2[56] and Minimap2[57] respectively. Stringtie version 1.3.6 was used to assemble transcripts from Ribo-seq data, with

reads mapping to each strand processed separately.[58] Pinfish was used to assemble transcripts from long reads (Oxford Nanopore Technologies). After transcript assembly, PASA[59] was used to combine the Stringtie and Pinfish models into a single transcriptome. Transdecoder[60] was then run twice: first, to identify candidate coding regions with PASA model with a lower limit of 100 amino acids, and second, to identify coding regions in just the Stringtie model with a lower limit of 40 amino acids. The latter run has a reduced risk of misannotating start codons in the 5′-UTR. Transdecoder annotated transcripts from $\text{Transdecoder}_{\text{PASA}}$ were used to train GlimmerHMM[61] and CodingQuarry,[62] which were used to provide de novo predictions in the genome. EVidenceModeler[63] was used to incorporate predictions from PASA, $\text{Transdecoder}_{\text{Stringtie}}$, $\text{Transdecoder}_{\text{PASA}}$, GlimmerHMM and CodingQuarry. File processing, UTRs, and tRNAs annotations were provided by the update utility in the Funannotate package.[64]

1.3.5 Mapping of ribosome protected reads to codons and masking

Linker ligated sequences were trimmed and demultiplexed in an error-tolerant way using Cutadapt.[65] Ribo-seq reads were mapped to the genome of *Komagataella pastoris* GS115[46] using HISAT2.[56,66] Alignments were converted from SAM to sorted and indexed BAM files using Samtools and only included reads with mapping quality threshold of 60.[67] Mapped reads were loaded into R using the GenomicAlignments package from Bioconductor[68] and converted to their 3′ end positions before determining p-site offsets. P-site offsets were determined using the RiboProfiling package in Bioconductor.[69] Each read was mapped to a single codon. Masking files were created by first parsing the coding sequence (CDS) 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 reference genome twice, one 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 reads per codon per gene (RPCPG) data tables. 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 (ORFs) were masked to correct for variable read quality at the beginning and ending of transcripts inherent to Ribo-seq.[70]

1.3.6 Metagene correction and quantification of metabolic demand

Read counts were normalized at the codon level using a metagene analysis that provides a global profile for each data set. First, for each ORF, reads at each codon position were scaled by the average reads per codon mapped ORF. Then, for codon position, either a mean or median value was calculated from all ORFs using the following scheme: for positions 1 to 100, a rolling mean with a window of 10 codons; for positions 100 to 1000, a rolling mean with a window of 100; for positions 1000 and onward, a rolling median with a window of 1000. In calculating corrected transcripts per million (cTPM), codon read counts were scaled by dividing the metagene-derived value at that position and normalized by their pseudo gene lengths (theoretical gene length minus number of masked codons) and a per million scaling factor unique to each data set. In calculating ribosomes per million (cRPM), a ribosome scaling factor was created for each gene by dividing the sum of the metagene-derived values at all codon positions by the sum of smoothed reads per codon with the mask applied (a gene with zero masked codons will have a ribosome

scaling factor equal to one, while a gene that contains masked codons will have a scaling factor greater than one). The ribosome scaling factor is multiplied by unmasked gene read counts and normalized by a per million scaling factor unique to each data set to give RPM. Membrane enrichment is quantified for each gene as the log_2 ratio of membrane cTPM scores or total cTPM scores to soluble cTPM scores.

1.3.7 Classification of ORFs

Gene names were hierarchically assigned to novel *K. phaffii* transcripts through homology. Firstly, transcripts were assigned names inherited from *S. cerevisiae*[71] using BlastP[72] with an expected value less than 1e-5. For genes that were not predicted to be homologous, gene names were assigned common names using EggNOG 4.5[73] using a taxonomic scope limited to ascomycetes. Genes that did not share homology with *S. cerevisiae* or known ascomycetes were assigned names inherited from *K. phaffii* GS115[46] using BlastP with expected values less than 1e-5. Novel genes that were not assigned names using the methods above were named after the moniker automatically assigned during transcript assembly.

ORFs were classified by function, cellular location, and sequence features using various prediction softwares. Functions were assigned ontologically using clusters of orthologous groups (COG) and were prepared using EggNOG 4.5.[73] Vironoi tessellations were created to quantitatively map the biosynthetic composition of these functions using COGs and expression metrics derived from Ribo-seq.[74] DeepLoc was used to predict the subcellular localization associated with ORF products.[75] Sequence features such as signal sequences, transmembrane domains (TMD), and GPI anchors were identified using SignalP 5.0,[76] TOPCONS,[77] and predGPI[78] re-

spectively.

1.4 Perspectives

1.4.1 Ribo-seq and long-read RNA-seq improve open reading frame annotations

We sought to globally quantify several aspects of protein synthesis in *K. phaffii* GS115. We asked which genes were responsible for sequestering limited biosynthetic resources, such as ribosomes and ER translocons. We also asked which genes were responsible for producing the most nascent chains, which is critical for predicting amino acid usage, as well as modifications that act on a per chain basis (i.e., N-terminal acetylation, GPI anchoring, vesicular sorting). Ribo-seq provides a snapshot of protein translation, allowing us to answer both of these questions.[33] It is a high throughput sequencing technique that provides a snapshot of protein translation and is capable of inferring ribosome abundance at each codon of each transcript.[79] In Ribo-seq, a non-specific ribonuclease generates $20\,\mathrm{nt}$ to $22\,\mathrm{nt}$ or $28\,\mathrm{nt}$ to $30\,\mathrm{nt}$ "footprints" of ribosome-protected mRNA depending on the translation conformation of the ribosome,[33] which are then sequenced. Ribo-seq libraries revealed adequate depletion of ribosome derived RNA using rRNA depletion strategies designed for *S. cerevisiae* in YPD media. Our data sets captured footprint lengths from $15\,\mathrm{nt}$ to $42\,\mathrm{nt}$(*Figure 1.7a*). Nearly all ($99\,\%$) footprints mapped within open reading frames (ORFs). Our profiling data also indicate active translation through the appearance of three nucleotide periodicity in read depth that is preserved across the transcriptome (*Figure 1.7b*).

We noticed that ribosome-protected read patterns were often inconsistent with

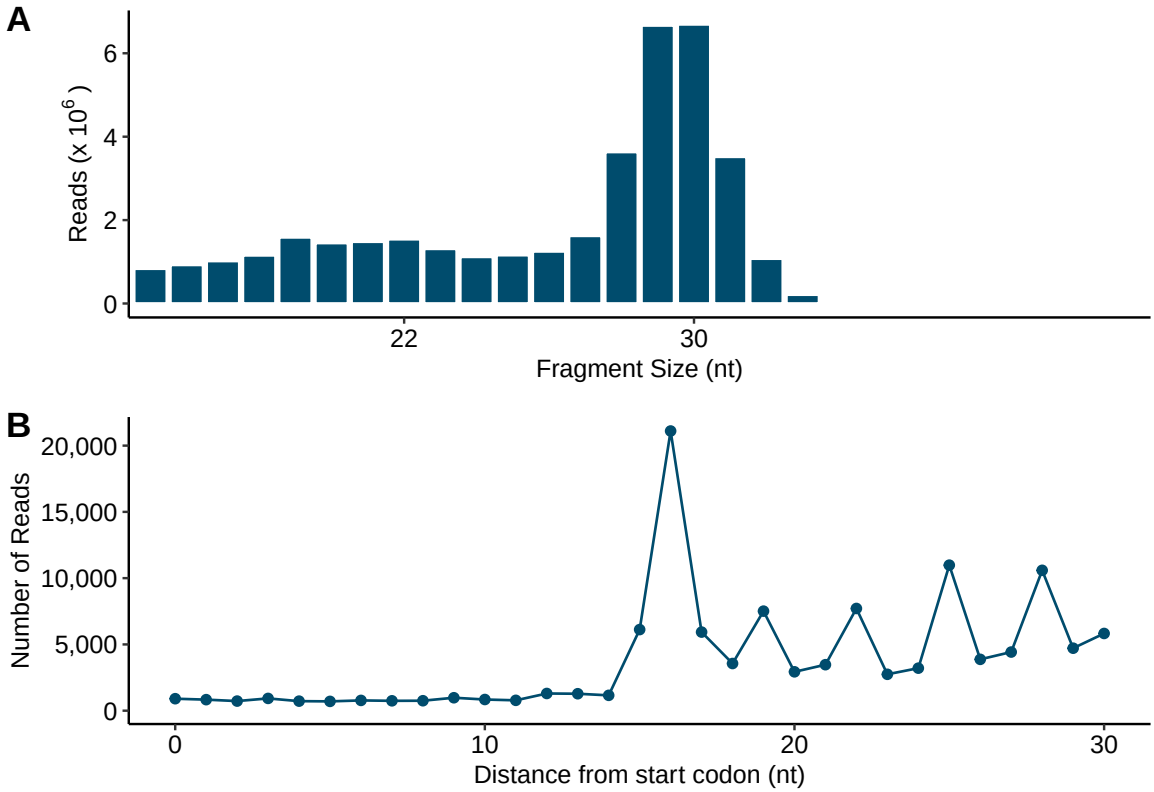

Figure 1.7: **Ribo-seq models active translation** ***a.*** Distribution of reads for different length RNA fragments. Read distribution is bimodal and reveals two peaks at 22 nt and 30 nt. These peaks reflect the distance that ribosomes cover mRNA and is dependent on stage of translocation. ***b.*** Nucleotide periodicity for 30nt fragment reveals active translation. The distance from the beginning codon to the codon where periodicity begins is known as a p-site offset. Calculating this offset helps to accurately map reads to codons.

prior annotations of open reading frames (*Figure 1.8*). At many loci, Ribo-seq appeared to indicate that translation began at an alternate start codon. Inaccuracies in ORF structure are problematic, since the length of a reading frame is a critical parameter used for quantifying translation and the position of the start site is used in correction using global profiles. We therefor sought to improve the GS115 annotation using Ribo-seq. Several methods that rely solely on Ribo-seq to annotate structure rely on the three nucleotide periodicity of reads to define reading frames.[80] They require substantial coverage for each genes; however, sparse Ribo-seq coverage could still support re-annotation if it were treated like stranded RNA-seq data.

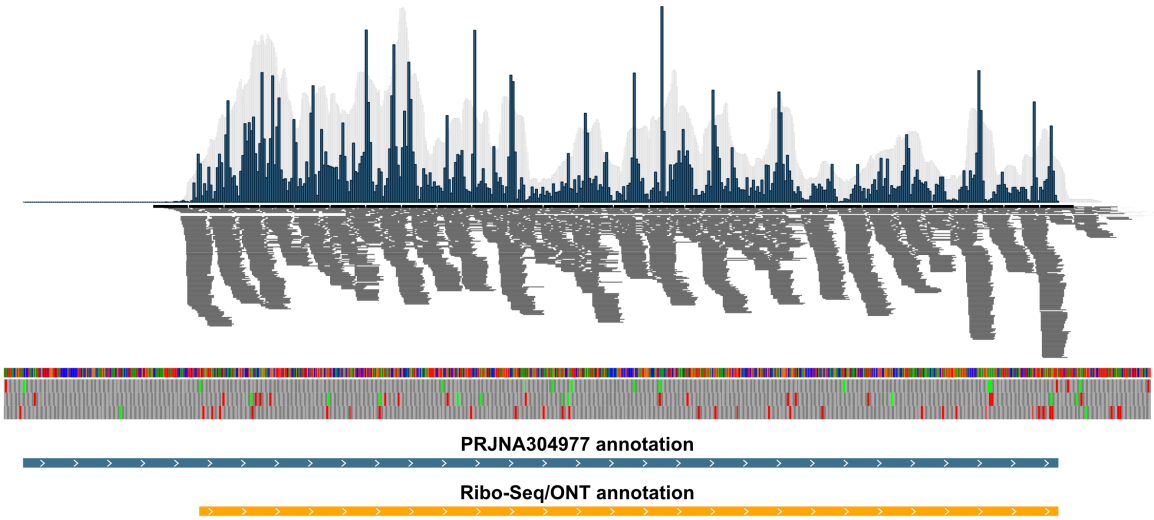

Figure 1.8: **Ribo-seq and long read RNA sequencing improves protein predictions** Images are adaptations from screen captures using Integrated Genome Viewer (MIT) and reflect reads for the gene *TIF1*. Ribosome-protected footprint reads mapped to transcripts are translighted left to right in gray stacks. Reads mapped to transcripts in R are shown as blue histograms while IGV's predicted read mapping are shown behind as light gray. Translation table for genome annotation are shown as multicolored panels where green squares represent theoretical start sites and red squares represent theoretical stop sites. Previous annotation incorrectly predicts start site and is shown in blue while Ribo-seq/ONT annotation correctly predicts start site and is shown in yellow.

Moreover, de novo open reading frame predictors can be trained using verified translational start sites, and so improving the accuracy of annotations for a subset of the transcriptome was expected to improve overall prediction accuracy. We therefore adapted consensus methods used in gene prediction and annotation with stranded RNA-seq data, with optimizations for fungi.[63,64] Our approach uses Ribo-seq to construct transcript models, which are then used to train several de novo annotators.

Like other yeasts, *K. phaffii* has short intergenic sequences, leading to overlapping untranslated regions (UTRs), even on transcripts encoded on the same DNA strand. As a results, methods that construct transcripts from short-read sequencing merge data from adjacent genes into a single transcript. We therefore collected long-read data using Oxford Nanopore PCR-cDNA sequencing and developed a

pipeline to integrate Ribo-seq, long-read RNA-seq, and de novo gene prediction (*Figure 1.8* and *Figure 1.9*). ORFs that were fully covered by Ribo-seq data were

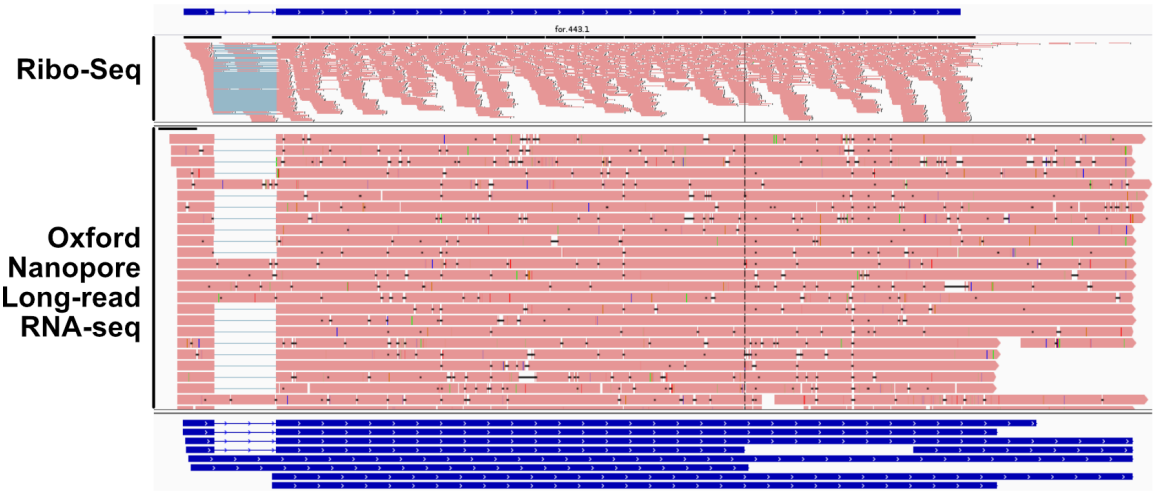

Figure 1.9: **Ribo-seq and long-read RNA-seq improve transcriptome annotation** Images are screen captures from Integrated Genome Viewer (MIT). In an example transcript, Ribo-seq (top register) and long-read RNA-seq (bottom register) reveal both the open reading frame and the untranslated regions (UTRs).

allowed to be as short as 40 amino acids, increasing the number of annotated genes compared to other annotations of *K. phaffii* (*Table 1.4*).[46,81,82] Our annotation adjusted the translational start site of about 10 % of ORFs compared to each previous model. Overall, Ribo-seq reads were mapped to 5,303 genes in *K. phaffii* in the assembly presented here. We have named genes based on homology to prior annotations, to *S. cerevisiae*, and to other ascomycetes.

Table 1.4: Comparing annotations

Annotation[a]	Total sequences	Homologs[b]	Length differences[c]
Ribo-seq annotation	5,329		
GS115 (PRJNA304976)	5,064	5,035	514
GS115 (PRJEA37871)	5,040	5,100	697
CBS7435 (PRJEA62483)	5,291	5,198	604

[a] NCBI bioproject numbers located in parenthesis
[b] BlastP matches from current study to prior study
[c] Number of homologs with different predicted lengths

1.4.2 Quantification of protein synthesis demands

Each read in Ribo-seq originates from a translating ribosome. To quantify protein synthesis demands, the number of nascent polypeptide chains produced per unit time can be approximated using a modified form of the transcripts per million (TPM) metric used in RNA-seq. TPM has advantages over other metrics (RPKM or FPKM) for its intuitive interpretation during differential analysis and for its congruence with proteomics.[83,84] In RNA-seq, reads are generally long enough to be unambiguously mapped to the transcriptome, and they can be assumed to equally cover a transcript. In Ribo-seq, however, these assumptions do not hold, and biases due to ambiguous mapping must be corrected. Ribosome protected fragments are small, $22\,\mathrm{nt}$ to $30\,\mathrm{nt}$, and have the capacity to map to multiple mRNA sequences when the transcriptome contains homologous stretches. Ambiguously mapped reads can be handled in one of several ways, often with shortcomings. Discarding multi-mapped reads[85–88] depreciates read counts for highly expressed genes. Randomly assigning reads to ORFs with equivalent percentage of alignment[79,89,90] overestimates read counts for lowly expressed genes. Here, we adapt the method of Taggart et al.,[35] who used computational masks to exclude homologous segments of the predicted transcriptome. We calculated a mask of the *K. phaffii* transcriptome accounting for all possible $28\,\mathrm{nt}$ reads, excluding $3\,\%$ of codon positions available. To estimate gene expression via transcripts per million (TPM), reads must be scaled by ORF length. Unlike discarding or randomly assigning reads, masking adjusts the gene length to reflect mRNA positions available for analysis. However, masking alone is insufficient because ribosome protected reads are not evenly distributed across transcripts.

Ribosome-protected reads are more abundant near the 5′ end of ORFs.[79,91] This

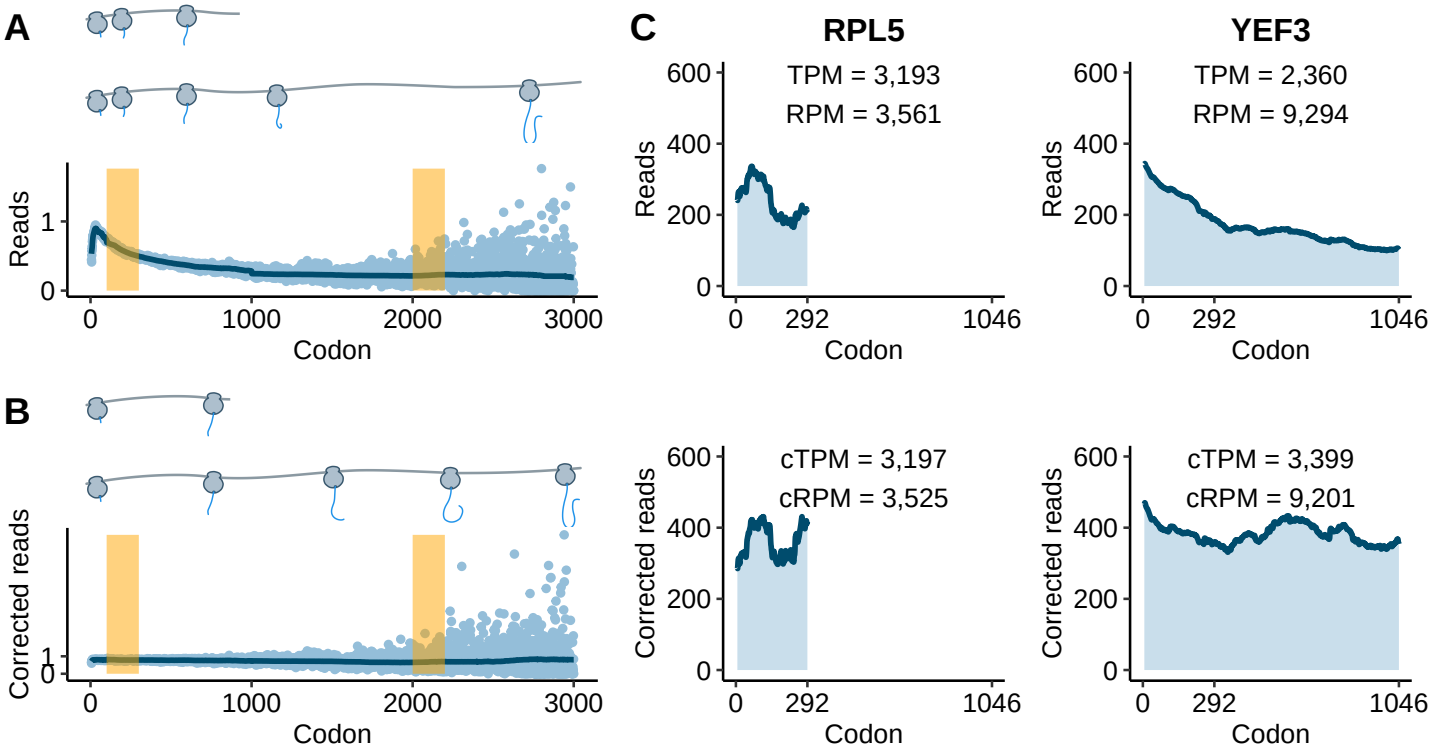

Figure 1.10: **Metagene analysis** Corrections applied to Ribo-seq data. ***a*** Ribosome-protected read counts at each codon were scaled by the total reads mapping to the ORF. Dots represent individual codons, and the line represents a composite of rolling means and medians (see Methods). Regions in yellow are the same width and are used to demonstrate that masked codons at the beginning of ORFs have a greater influence of calculated expression than masked codons at the end of ORFs. ***b*** Data from a after metagene correction. ***c*** Comparison of ribosome-protected reads per codon for highly expressed genes of different length. TPM for *RPL5* gene is approximately 135% greater than TPM for *YEF3* while producing approximately 38% as many ribosome-protected reads. After metagene correction cTPM scores are similar preserving the same difference in ribosome sequestration

effect may be due to slower elongation rates at the beginning of translation[92] or abortive translation.[35] Regardless of the mechanism, the positional bias is observed in nearly every transcript and results in a global read profile that is conserved across the translatome (*Figure 1.10a*). As a result, estimates of the expression of short ORFs will appear inflated (and long ORFs deflated), since only the ribosome-rich region of the global profile is sampled. We again adapt the method of Taggart et al.,[35] where the positional bias is removed by scaling reads at each codon by the empirical global profile (*Figure 1.10b*). We used corrected TPM (cTPM), with masking and scaling, as a measure of the rate at which nascent chains are produced. For example, transcripts of *RPL5* and *YEF3* display similar numbers of ribosomes at the start of their ORFs (*Figure 1.10c*), suggesting similar initiation rates. However, because *YEF3* is a longer ORF, its standard TPM is smaller than the TPM of *RPL5*. Here, we assume that if *RPL5* were as long as *YEF3*, then its translation profile will be similar to the global profile, resulting in similar cTPM scores.

While cTPM estimates the number of nascent polypeptide chains, it does not inform us about ribosome sequestration. Longer transcripts sequester a greater number of ribosomes in order to produce the same number of nascent chains as a shorter transcript. If ribosomes accumulate near the start codon *in vivo*, then it is important to include this effect while measuring allocation. cTPM, therefore, is an appropriate metric. If ribosome-protected reads could be unambiguously mapped to the transcriptome, then simple read counts estimate ribosome usage per gene. However, when masking is applied, the position of the mask becomes important (*Figure 1.10a, b*). Two masks of the same length, applied at different positions, will hide different amounts of ribosomes based on the global profile. To correct for this, we introduce a ribosome scaling factor that accounts for masking of each gene. The factor represents the fraction of ribosomes expected to be observed when the

Table 1.5: **Nascent chains produced in *K. phaffii***

	Genes	Nascent chains[a]
Ontological functions		
Translation, ribosomal structure and biogenesis	366	44
Function unknown	1,602	11
Post-translational modification, protein turnover and chaperones	409	9
Energy production and conversion	207	8
Intracellular trafficking, secretion and vesicular transport	382	4
Carbohydrate transport and metabolism	218	3
Cell wall/membrane/envelope biogenesis	85	3
Amino acid transport and metabolism	191	3
Transcription	355	2
Rna processing and modification	242	2
Predicted features		
Lumenal and secreted proteins[b]	266	8
Gpi anchors[c]	117	79
Transmembrane proteins[d]	960	7

[a] Nascent chains are percentage of the total cTPM represented by each category
[b] Total number of genes with an N-terminal signal sequence and may include a GPI anchor
[c] Percentage of signal sequences that also contain a predicted GPI anchor
[d] Contain no signal sequence but one transmembrane domain (TMD), or two or more TMDs

gene-specific mask is applied to the global translational profile. We generate a new metric for each gene, correct ribosomes per million (cRPM), which is practically equivalent to reads per million (RPM) in standard RNA-seq. In our example in (*Figure 1.10c*), cRPM and RPM are almost identical, as expected since there are no masks applied to *RPL5* or *YEF3*.

After applying corrections, we find that the majority of nascent chains synthesized by *K. phaffii* are from genes involved in translation, ribosomal structure, and biogenesis (*Table 1.5* and *Figure 1.11*), as expected for log-phase growth. The majority of nascent chains encoded by genes of unknown function are predicted to be extracellular, where they are likely components of the cell wall. We consider endomembrane lumenal and secreted proteins to be those with (i) predicted N-terminal signal sequences, (ii) are not predicted to be localized to the mitochondria, and (iii) contain less than or equal to one transmembrane domain, as these are

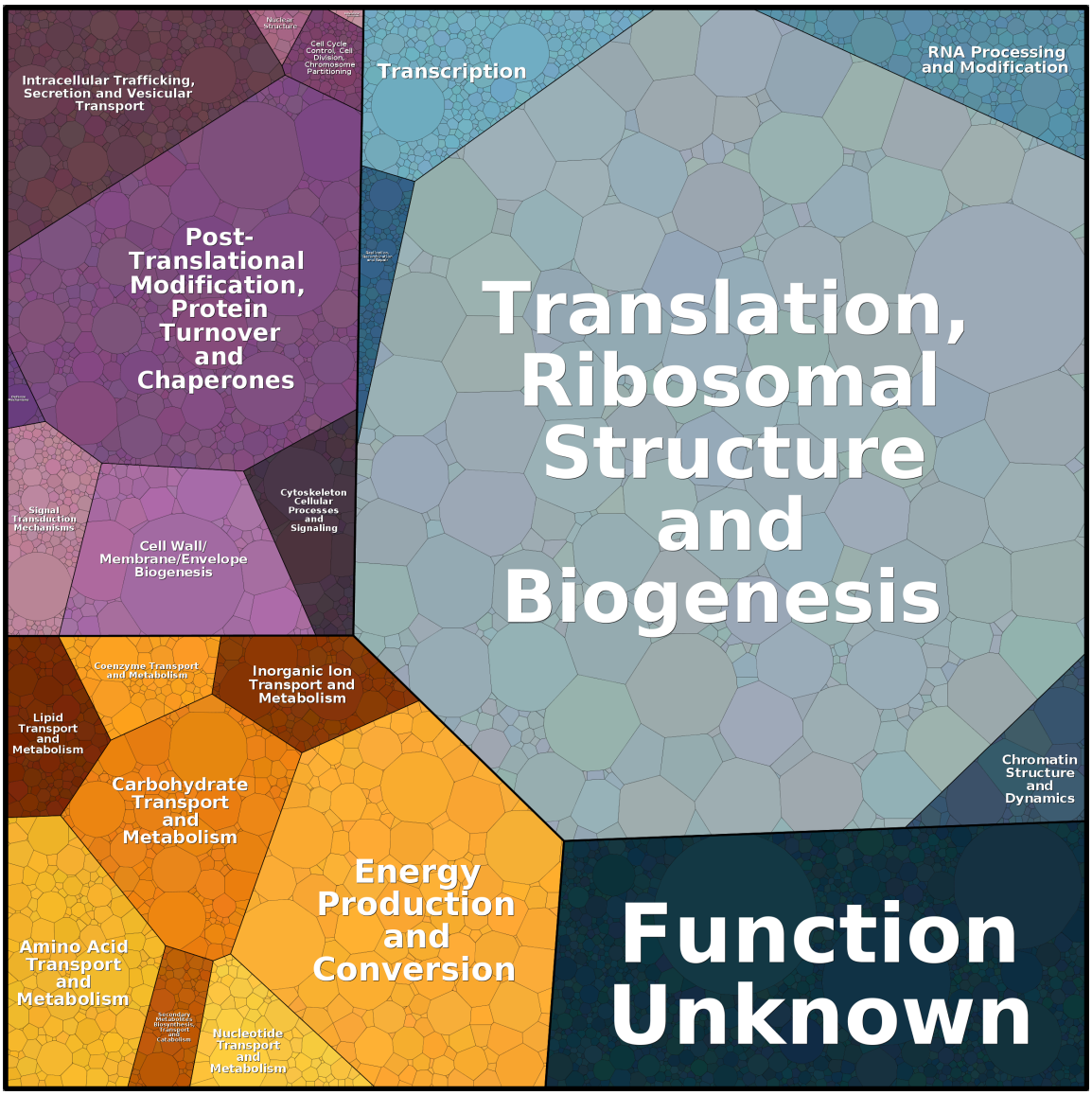

Figure 1.11: **Translational landscape in *K. phaffii*** Tessellations represent the relative number of nascent chains produced by ribosomes for each gene. Each gene is categorized by their predicted ontological function. A gene's tile size is proportional to their respective cTPM score within a Ribo-seq sample.

frequently GPI anchors. Some single-pass, type I transmembrane proteins will be misannotated by this definition. The number of genes containing these predictive features and the relative percentage of nascent chains they produce are summarized in *Table 1.5*. A majority of nascent chains for genes containing a signal sequence also

contain GPI anchors, suggesting that this structural class represents the majority of products that will be processed by the secretory pathway.

1.5 Conclusions

Ribo-seq is an enormously powerful technology that allows us to detect and quantify protein translation *in vivo*. This technology allows for the quantification of metabolic demands, understanding of translational control mechanisms, and the determination of encoded transcript regions for a more thorough understanding of the translatome. While protocols for Ribo-seq and its analysis are ever changing and under constant development, we offer a standardized computational pipeline. First, we show that a commercially available rRNA depletion strategy designed for the model organism *S. cerevisiae* is effective in *K. phaffii* collected in YPD during exponential growth. Second, we provide an updated genome annotation based on both Ribo-seq and long-read RNA-seq. Third, we address the inherent propensity for small reads to map to multiple locations in the genome using a computational masking strategy. Finally, we utilize this protocol to broadly characterize nascent chain production for genes characterized by their ontological functions.

Chapter 2

Characterization of endoplasmic reticulum translocation pathways and comparison of early secretory demands in *Komagataella phaffii* and *Saccharomyces cerevisiae*

Background: Eukaryotes use distinct networks of biogenesis factors to synthesize, fold, monitor, traffic, and secrete proteins. During heterologous expression, saturation of any of these networks may bottleneck titer and yield. However, most of what we know about these processes is derived from *Saccharomyces cerevisiae*. To understand the flux through various routes into the early secretory pathway, we quantified the global and membrane-associate translatomes of *Komagataella phaffii*.

Results: By using ribosome profiling with subcellular fractionation, we quantified demands on co- and post-translational translocation pathways. During ex-

ponential growth in rich media, protein components of the cell-wall represent the greatest number of nascent chains entering the ER. Transcripts encoding the transmembrane protein *PMA1* sequester more ribosomes at the ER membrane than any others. Comparison to *Saccharomyces cerevisiae* reveals conservation in the resources allocated by gene ontology, but variation in the diversity of gene products entering the secretory pathway.

Conclusion: A subset of host proteins, particularly cell-wall components, impost the greatest biosynthetic demands in the early secretory pathway. These proteins are potential targets in strain engineering aimed at alleviating bottlenecks during heterolgous protein production.

2.1 Introduction

Identifying and relieving protein biogenesis bottlenecks is one strategy to improve yields of high-value, recombinant proteins.[2,93] For secreted proteins expressed in *K. phaffii*, an early bottleneck is the translocation of newly made proteins from the cytoplasm into the lumen of the endoplasmic reticulum (ER).[19,20] Yeasts have multiple pathways for translocation, which use partially overlapping sets of biogenesis factors (reviewed in).[3] In the major pathway into the ER, translocation occurs through a membrane-embedded protein complex called the sec translocon. At least three major translocons exist in yeasts (the Ssh1 complex; two Sec61 complexes with, and without, Sec62p, Sec63p, Sec66p and Sec71p), which can accept proteins as they are synthesized by ribosomes (co-translationally) or after synthesis of the polypeptide chain is complete (post-translationally). Besides translocon architecture, co- and post-translational pathways differ in their reliance on cytosolic molecular chaperones.[94,95] Translocons bind hydrophobic amino acid motifs, called signal

peptides, found at the amino termini of secreted proteins.[96] Some signal peptides are dependent upon a cytosolic factor, the Signal Recognition Particle (SRP), and the ER-bound SRP receptor to engage a translocon;[97] these tend to be longer or more hydrophobic than SRP independent signals.[98,99] Binding of a signal peptide to a translocon opens the channel and allows the rest of the protein to pass into the lumen. In addition to secreted proteins, the sec translocon is a major point of entry for integral membrane proteins of the endomembrane system.[100] Integral membrane proteins that use a sec translocon require SRP for targeting to the ER over mitochondria.[98]

For any production host, ribosomes, molecular chaperones, and sec translocons represent limited pools of resources that are distributed between heterologous proteins and the host proteome.[103] Unlike resources that are replenished enzymatically (like aminoacyl-tRNAs), ribosomes, translocons and chaperones only act on a single nascent chain at a time. While in use, they are sequestered and unavailable for other tasks. Although computational models that approximate these effects exist for bacteria,[104] the complexity of eukaryotic translation is insufficiently understood to predict these allocations from transcriptomics alone. Accurate accounting of these resources could allow strains to be engineered in ways to relieve bottlenecks specific to a target. The secretome of *K. phaffii* has been characterized under several conditions,[105] but the precise biosynthetic requirements of each protein remain unknown. Sequence features of secreted proteins, like glycosylation motifs, allow approximation of their direct biosynthetic costs such as ATP, carbohydrates, disulfide bonds, or GPI-anchors.[106] Per molecule costs can be coupled with measurements of gene expression to identify most expensive host proteins. Deletion of these proteins improves yields of secreted heterologous proteins in mammalian systems.[107,108] However, while these analyses account for demands on

global resources, they are limited by insufficient experimental data which links gene products to specific biogenesis subnetworks. For instance, overloading co-translational translocons could limit secretory yields even if metabolic demands are met and post-translational translocons are available. Quantification of global ribosome, co-translational translocon and SRP use is available for *S. cerevisiae*.[21,23,98] However, these measurements are unavailable for other industrially significant species, including *K. phaffii*.

Which host proteins sequester the most biogenesis machinery in the early secretory pathway of *K. phaffii*? Which host genes produce the most nascent chains, competing for chaperones and sorting factors within the endomembrane system? To answer these questions, we quantified active translation globally and at the surface of the ER or mitochondria using ribosome profiling (Ribo-seq). Our analysis reveals the set of proteins that enter the secretory pathway co-translationally and predicts the set that enter post-translationally. In each set, we estimate demand for ribosomes and translocons. We distinguish between resources that act on a per nascent chain basis from machinery that is utilized based on elongation time.

2.2 Materials and Methods

2.2.1 Strains and culture conditions

All experiments were performed us *Komagataella phaffii* GS115 (Invitrogen). For each Ribo-seq biological replicate, $500\,\mathrm{mL}$ liquid cultures of YPD ($1\,\%$ yeast extract, $2\,\%$ peptone and $2\,\%$ glucose) were grown to an $\mathrm{OD}_{600\,\mathrm{nm}}$ of 2 at $30\,^{\circ}\mathrm{C}$ with shaking in baffled $2\,\mathrm{L}$ flasks. Cells were harvested by vacuum filtration through a $0.8\,\mu\mathrm{m}$ filter. Immediately after filtering, cells were scraped off the filter using a chilled

scoopula and submerged in a $50\,\mathrm{mL}$ conical tube containing liquid nitrogen. When indicated in order to match conditions of *S. cerevisiae* fractionated Ribo-seq data,[21] cycloheximide (CHX) was added to $100\,\mu\mathrm{g\,mL^{-1}}$ for $3\,\mathrm{min}$ prior to harvesting. CHX treatments longer than a few minutes can alter ribosome abundance near the start of transcripts.[109] Short incubation with CHX enhance targeting of translocation competent ribosome-nascent chain complex while not perturbing non-secretory polysomes.[23]

2.2.2 Lysis and subcellular fractionation

Cells were lysed in either soluble lysis buffer ($50\,\mathrm{mM}$ MOPS, $25\,\mathrm{mM}$ KOH, $100\,\mathrm{mM}$ KOAc, $2\,\mathrm{mM}$ MgOAc, $1\,\mathrm{mM}$ DTT and $100\,\mu\mathrm{g\,mL^{-1}}$ CHX) or membrane lysis buffer (soluble lysis buffer with $1\,\%$ Triton X-100). Lysis buffers for each sample were frozen by adding $2\,\mathrm{mL}$ per dropwise to a $50\,\mathrm{mL}$ conical tube containing liquid nitrogen. For each biological replicate, $\frac{2}{3}$ frozen cells were mixed with $2\,\mathrm{mL}$ frozen soluble lysis and the remaining $\frac{1}{3}$ were mixed with $2\,\mathrm{mL}$ frozen membrane lysis buffer. Cell fractions were pulverized for $2\,\mathrm{min}$ in a $50\,\mathrm{mL}$ ball mill chamber with a single $2\,\mathrm{mL}$ steel ball (Retsch) collected into $1.5\,\mathrm{mL}$ conical tubes. After thawing, lysates were centrifuged at $20\,000\,\mathrm{g}$ for $10\,\mathrm{min}$. Supernatants from samples lysed with membrane lysis buffer were collected and used as "total" fractions. Supernatants from samples lysed with soluble lysis buffer were collected and used as "soluble" fractions. The pellets from sample lysed with soluble lysis buffer were resuspended in $2\,\mathrm{mL}$ membrane lysis buffer and centrifuged. The supernatants were collected and used as "membrane" fractions. Triton-X 100 was added to $1\,\%$ in soluble fractions, so that all three fractions were in equivalent buffers (*Figure 2.1*).

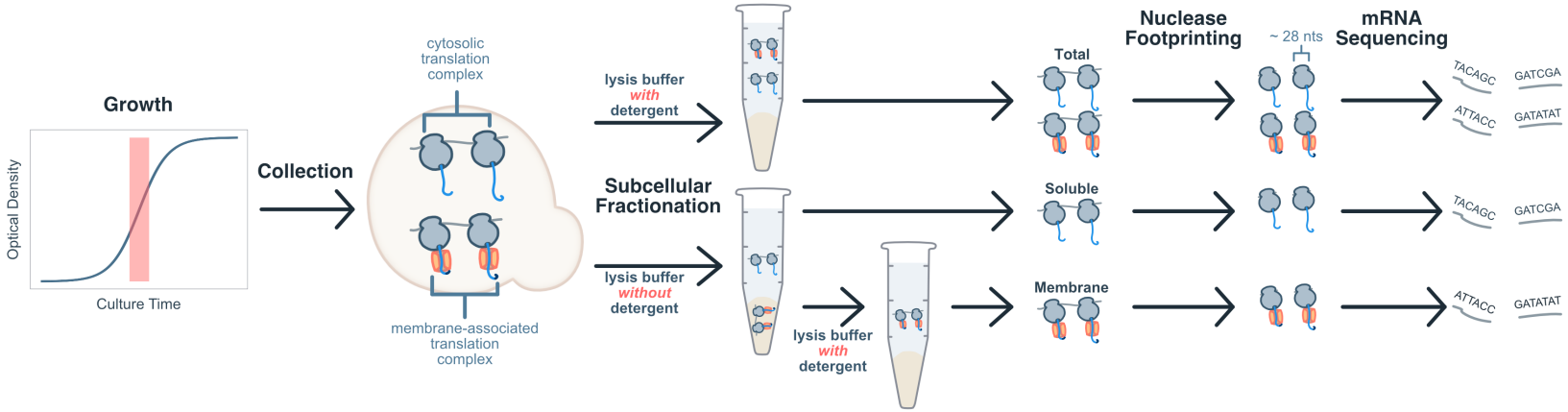

Figure 2.1: **Overview of Ribo-seq and subcellular fractionation** Ribosomes (gray) bound to a translocon (red) are only solubilized in the presence of detergent. The total sample has footprints originating from both membrane-bound and free-floating ribosomes. The soluble fraction is enriched in footprints from free-floating ribosomes. The membrane fraction is enriched in footprints from membrane-bound ribosomes

2.2.3 Ribo-seq

Lysed samples were nuclease digested using $40\,\mathrm{U}$ of Ambion RNase A for $1\,\mathrm{h}$ at room temperature. Digested samples were layered on a $10\,\%$ to $50\,\%$ sucrose gradient prepared in $50\,\mathrm{m}$M Tris pH 7.5, $200\,\mathrm{m}$M NaCl, and $2\,\mathrm{m}$M MgOAc case using a Gradient Master (Biocomp). Gradients were centrifuged at $39\,000$ RPM for $2.5\,\mathrm{h}$ 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 $18\,\mathrm{nt}$ to $34\,\mathrm{nt}$ were resolved and excised using $15\,\%$ polyacrylamide TBE-urea gel. RNA was collected from excised gel fragments using RNA gel extraction buffer ($300\,\mathrm{m}$M NaOAc, $1\,\mathrm{m}$M EDTA, and $0.25\,\%$ SDS), precipitated, and resuspended in water containing $20\,\mathrm{U\,mL^{-1}}$ SUPERase · In.

Purified fragments were then dephosphorylated by incubating $2\,\mu\mathrm{L}$ $1\,\mathrm{M}$ RNA sample with $2\,\mu\mathrm{L}$ RNase free water, $0.5\,\mu\mathrm{L}$ SUPERase · In RNase Inhibitor, $0.5\,\mu\mathrm{L}$ T4 Polynucleotide Reaction Buffer (PNK) (NEB, Cat #B0201S), and $0.5\,\mu\mathrm{L}$ T4 Polynucleotide Kinase at $37\,^{\circ}\mathrm{C}$ for $1\,\mathrm{h}$. Dephosphorylated samples were linker ligated with adapter sequences by incubating with $3.5\,\mu\mathrm{L}$ $50\,\%$ PEG-8000, $0.5\,\mu\mathrm{L}$ 10X T4 RNA Ligase Reaction Buffer, $0.5\,\mu\mathrm{L}$ $10\,\mu$M adenylated linkers and $0.5\,\mu\mathrm{L}$ T4 Rnl2(tr)k277Q at $30\,^{\circ}\mathrm{C}$ for $4\,\mathrm{h}$. Linker-ligated samples were concentrated via isopropranol 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 man-

ufacturer. Depleted samples were precipitated, resolved using $15\,\%$ TBE-urea polyacrylamide gel, and extracted as previously described.

RNA was reverse transcribed by adding $2\,\mu L$ reverse transcription primer to $10\,\mu L$ sample and incubating at $65\,°C$ for $5\,min$ to denature. Denatured sample was then incubated with $4\,\mu L$ 5X First Strand Buffer, $1\,\mu L$ $10\,mM$ dNTPs, $1\,\mu L$ $10\,mM$ DTT, $1\,\mu L$ $20\,U\,\mu L^{-1}$ SUPERase · In and $1\,\mu L$ $200\,U\,\mu L^{-1}$ SuperScript II Reverse Transcriptase at $50\,°C$ for $30\,min$ using thermal block. After incubation, sample was hydrolyzed by adding $2.2\,\mu L$ $1\,M$ NaOH and then incubated at $70\,°C$ for $20\,min$ using thermal block. $28\,\mu L$ RNAse free water was added to reverse transcription mixture (~$50\,\mu L$ 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 method previous described. Extracted precipitants were resuspended in $11\,\mu L$ 1:1000 SUPERase · In.

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

Circularized samples were PCR amplified for 12 cycles using a $50\,\mu L$ reaction mixture ($10\,\mu L$ Q5 Reaction Buffer (NEB , Cat #B9027S), $1\,\mu L$ $10\,mM$ dNTPs, $2.5\,\mu L$ $10\,\mu M$ forward primer, $4\,\mu L$ circularized DNA sample, $0.5\,\mu L$ Q5 High Fidelity DNA Polymerase and $29.5\,\mu L$ RNAse free water) divided into 5 x $10\,\mu L$ 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 HiSeq 4000.

2.2.4 Mapping of ribosome protected reads to codons and masking

Linker ligated sequences were trimmed and demultiplexed in an error-tolerant way using Cutadapt.[65] Ribo-seq reads were mapped to the genome of *Komagataella pastoris* GS115[46] using HISAT2.[56,66] Alignments were converted from SAM to sorted and indexed BAM files using Samtools and only included reads with mapping quality threshold of 60.[67] Mapped reads were loaded into R using the GenomicAlignments package from Bioconductor[68] and converted to their 3′ end positions before determining p-site offsets. P-site offsets were determined using the RiboProfiling package in Bioconductor.[69] Each read was mapped to a single codon. Masking files were created by first parsing the coding sequence (CDS) 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 reference genome twice, one 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 reads per codon per gene (RPCPG) data tables. 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 (ORFs) were masked to correct for variable read quality at the beginning and ending of transcripts inherent to Ribo-seq.[70]

2.2.5 Metagene correction and quantification of metabolic demand

Read counts were normalized at the codon level using a metagene analysis that provides a global profile for each data set. First, for each ORF, reads at each codon position were scaled by the average reads per codon mapped ORF. Then, for codon

position, either a mean or median value was calculated from all ORFs using the following scheme: for positions 1 to 100, a rolling mean with a window of 10 codons; for positions 100 to 1000, a rolling mean with a window of 100; for positions 1000 and onward, a rolling median with a window of 1000. In calculating corrected transcripts per million (cTPM), codon read counts were scaled by dividing the metagene-derived value at that position and normalized by their pseudo gene lengths (theoretical gene length minus number of masked codons) and a per million scaling factor unique to each data set. In calculating ribosomes per million (cRPM), a ribosome scaling factor was created for each gene by dividing the sum of the metagene-derived values at all codon positions by the sum of smoothed reads per codon with the mask applied (a gene with zero masked codons will have a ribosome scaling factor equal to one, while a gene that contains masked codons will have a scaling factor greater than one). The ribosome scaling factor is multiplied by unmasked gene read counts and normalized by a per million scaling factor unique to each data set to give RPM. Membrane enrichment is quantified for each gene as the log_2 ratio of membrane cTPM scores or total cTPM scores to soluble cTPM scores.

2.2.6 Classification of ORFs

Gene names were hierarchically assigned to novel *K. phaffii* transcripts through homology. Firstly, transcripts were assigned names inherited from *S. cerevisiae*[71] using BlastP[72] with an expected value less than 1e-5. For genes that were not predicted to be homologous, gene names were assigned common names using EggNOG 4.5[73] using a taxonomic scope limited to ascomycetes. Genes that did not share homology with *S. cerevisiae* or known ascomycetes were assigned names

inherited from *K. phaffii* GS115[46] using BlastP with expected values less than 1e-5. Novel genes that were not assigned names using the methods above were named after the moniker automatically assigned during transcript assembly.

ORFs were classified by function, cellular location, and sequence features using various prediction softwares. Functions were assigned ontologically using clusters of orthologous groups (COG) and were prepared using EggNOG 4.5.[73] Vironoi tessellations were created to quantitatively map the biosynthetic composition of these functions using COGs and expression metrics derived from Ribo-seq.[74] DeepLoc was used to predict the subcellular localization associated with ORF products.[75] Sequence features such as signal sequences, transmembrane domains (TMD), and GPI anchors were identified using SignalP 5.0,[76] TOPCONS,[77] and predGPI[78] respectively.

2.2.7 *S. cerevisiae* analysis

Ribo-seq data for total protein synthesis were taken from,[35] and data obtained from soluble or membrane-bound ribosome fractions were obtained from.[21] All data were processed in the same way as *K. phaffii* using the S288C reference genome R64-2-1.[110]

2.3 Results

2.3.1 Biogenesis demands in the early secretory pathway

We investigated the global demands for machinery needed for translocation into the ER. Subcellular fractionation was used to separate membrane-bound ribosomes from free floating, soluble ribosomes. Membrane-bound ribosomes were detergent

solubilized, and then samples from both soluble and membrane fractions were subject to Ribo-seq (*Figure 2.1*). As in *S. cerevisiae*, libraries derived from the membrane fractions are enriched in ribosome-protected footprints originating from transcripts that encode proteins destined for the ER or mitochondria.[21] Membrane enrichment scores were calculated as the log_2 ratio of cTPM for membrane and soluble fractions and were reproducible. The magnitude of membrane enrichment scores depends on the efficiency of fractionation, and if a gene falls below the diagonal line in *Figure 2.2*, it will have a negative enrichment score. As in *S. cerevisiae*, membrane enrichment

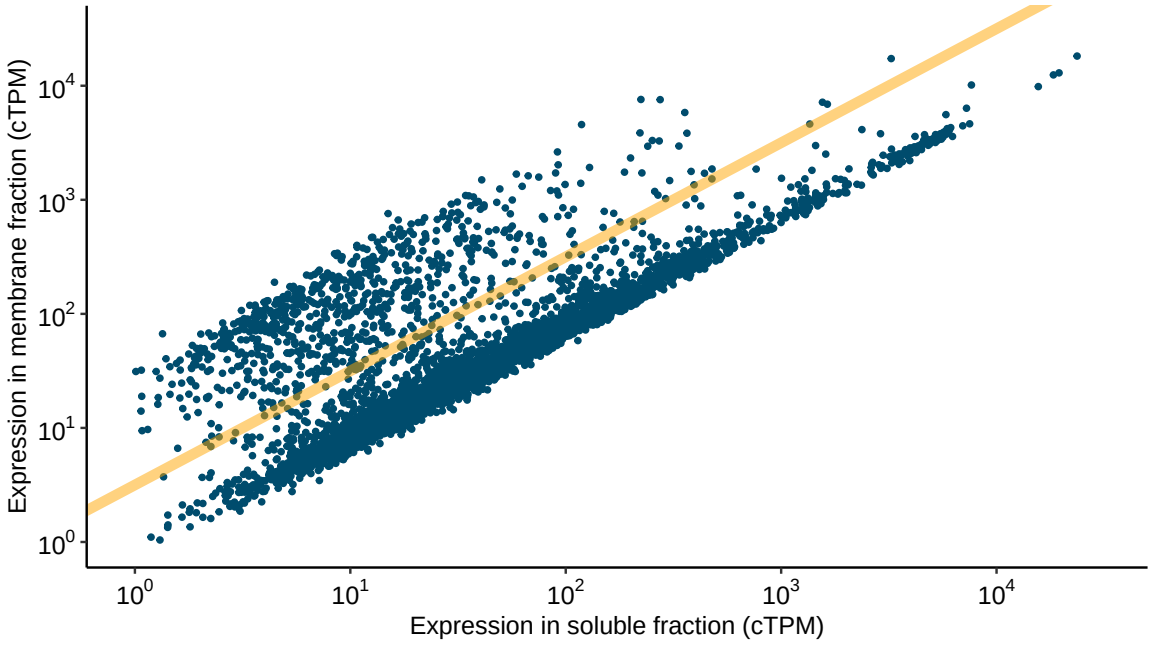

Figure 2.2: **Comparison of translation from samples of membrane-bound and soluble fraction** Values are calculated using fractions obtained after incubation with CHX. Genes that are considered membrane enriched have a log_2 ratio of expression on the membrane and cytosol. This cutoff is represented as a linear yellow line.

scores are limited by the length of the ORF when transcripts encode signal-sequence bearing proteins[21,23] (*Figure 2.3*). This effect is due to a kinetic competition between trafficking rate and translation elongation rate. *Figure 2.2* also reveals that a membrane enrichment score of 2 effectively separates two populations, and so

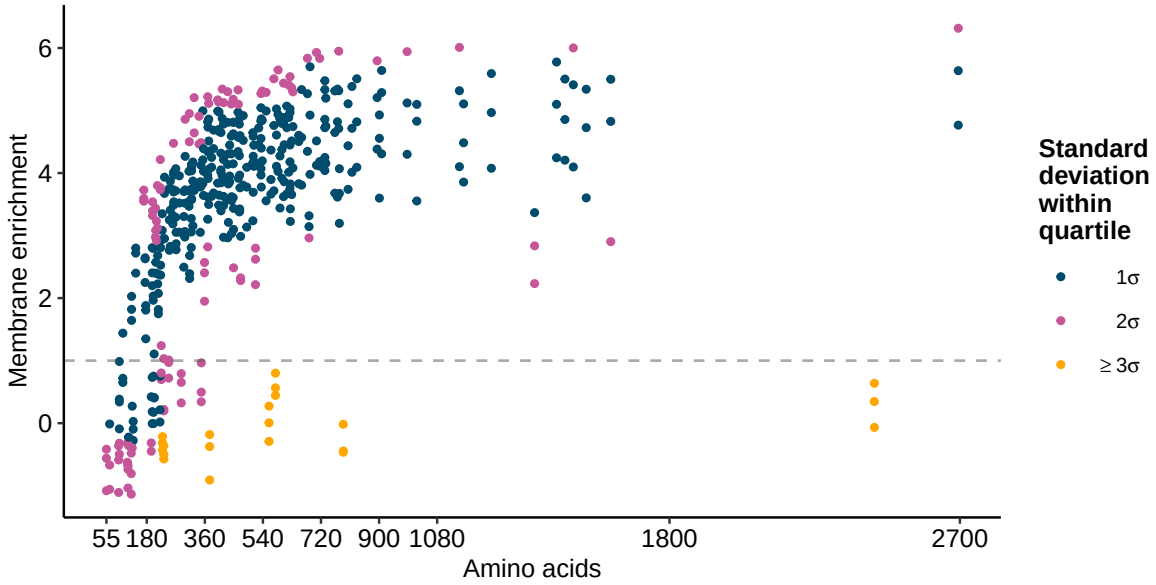

Figure 2.3: **Nascent peptide length and membrane enrichment for secreted, lumenal, or GPI-anchored proteins** Proteins have a predicted N-terminal signal sequence. Three sets of biological replicates were included for membrane and cytosolic fractionated libraries. The dashed line is drawn to establish cutoff for log_2 membrane enrichment. Proteins under this line are considered post-translationally targeted and genes over this line are considered co-translationally translocated. Proteins were binned according to their probabilistic distribution. Proteins were color coded to represent the number of standard deviations from the mean membrane enrichment score for the bin they belong to.

we define genes with scores greater than 2 as co-translationally translocated into either the ER or mitochondria. The set of co-translationally translocated nascent polypeptides is enriched for those involved in energy production and conversion, cell wall and membrane biogenesis, and various transporters. To assess entry into the ER, we filtered out transcripts encoding proteins predicted to localize in the mitochondria by DeepLoc (*Figure 2.4a*). Finally, we define proteins that enter the ER through a post-translational *sec* translocon as those having a predicted N-terminal signal sequence and less than two-fold membrane enrichment (*Figure 2.4b*). Post-translationally trafficked membrane proteins rely on other mechanisms, such as the GET pathway.[96]

A more diverse group of proteins enter the ER through co-translational translo-

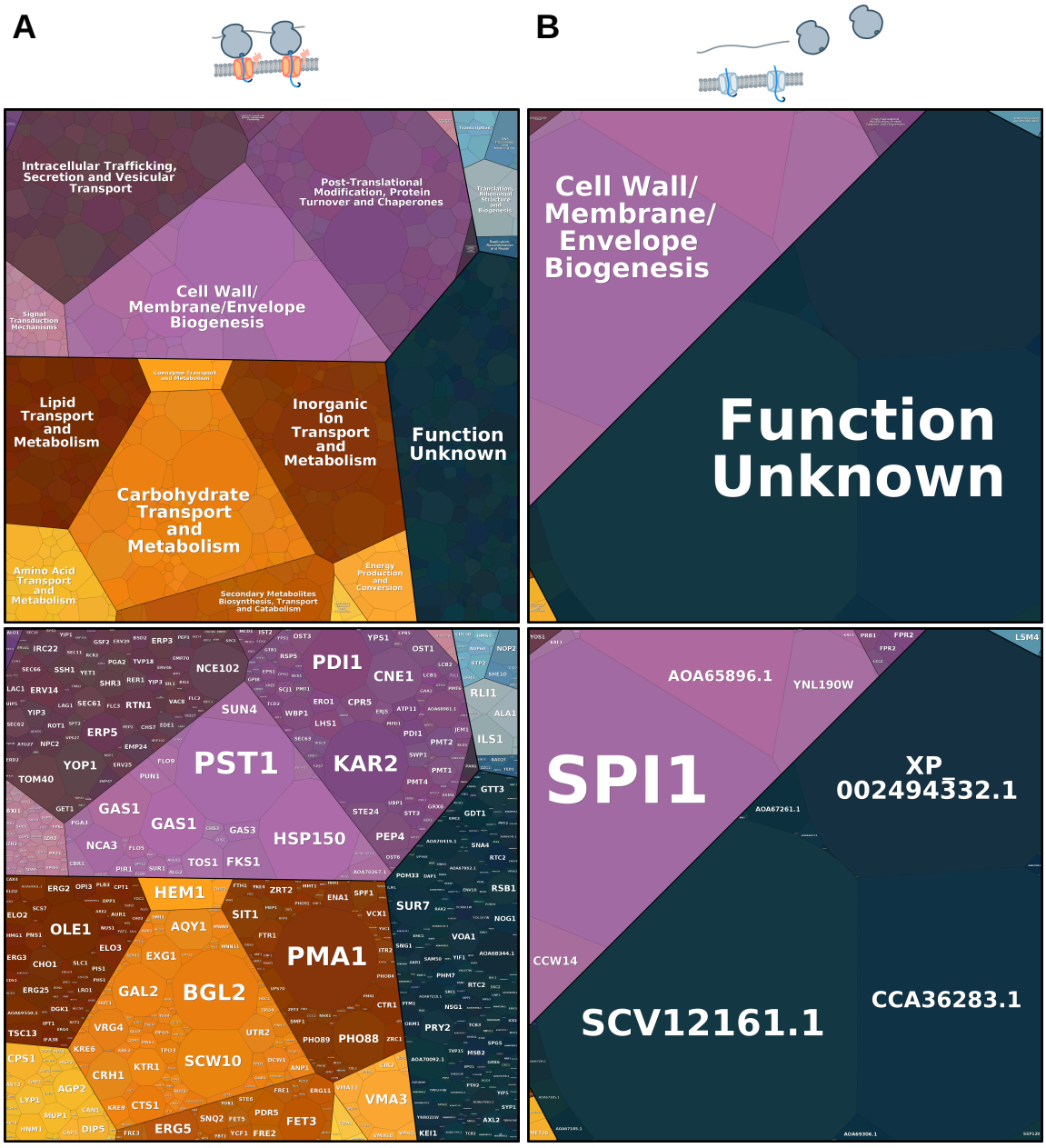

Figure 2.4: **Translation on the ER-membrane in *K. phaffii*** Tessellations are calculated using cTPM from the total fraction of a CHX treated culture and represent relative quantities of nascent chains produced by ribosomes associated with the ER-membrane for each gene. ***a.*** Translation of co-translationally translocated proteins. ***b.*** Translation of post-translationally translocated proteins.

cons than those that enter post-translationally (*Figure 2.4a,b* and *Table 2.1*). While the diversity of functions for proteins that enter the ER post-translationally is relatively small (mostly of unknown function and then cell wall and membrane

Table 2.1: **Comparison of translocon demands by ontological function**

Function	Genes	Nascent chains[a]	Ribosomes[b]
Co-translationally translocated[c]			
Function unknown	261	7.9%	10.7%
Cell wall/membrane/envelope biogenesis	41	7.4%	11.9%
Post-translational modification, protein turnover and chaperones	89	6.8%	11.6%
Carbohydrate transport and metabolism	114	6.6%	9.5%
Intracellular trafficking, secretion and vesicular transport	95	6.3%	7.4%
Inorganic ion transport and metabolism	82	4.9%	9.7%
Lipid transport and metabolism	72	3.8%	5.4%
Post-translationally translocated[d]			
Function unknown	30	33.4%	9.6%
Cell wall/membrane/envelope biogenesis	10	14.2%	8.9%
Post-translational modification, protein turnover and chaperones	5	0.3%	0.2%

[a] Calculated as percent of total cTPM for all proteins predicted to be ER destined
[b] Calculated as percent of total cRPM for all proteins predicted to be ER destined
[c] Greater than 2-fold membrane enrichment and not mitochondrial
[d] Lesser than 2-fold membrane enrichment, not mitochondrial and contain a signal sequence

biogenesis), we find that post-translational translocation handles a majority of total nascent chains entering the ER. These genes encode primarily small proteins such as Scv12161.1p or cell wall proteins processed with GPI-anchors, such as *SPI1*. Although its function is unknown, *SPI1* is also predicted to be GPI-anchored, and both *SPI1* and *SCV12161.1* produce among most nascent proteins within the cell under conditions tested here. We then classified the genes of unknown function that entered the ER by their predicted final location. The majority of these gene products, approximately four fifths, are predicted to be localized extracellularily and have an unusual discrepancy between their relative ribosomal usage, nascent chains produced, and average gene length compared to unknown genes predicted to localize elsewhere (*Table 2.2*).

Table 2.2: **Biosynthetic demands for proteins with unknown ontological functions by predicted subcellular localization**

Location	Genes	Mean length[a]	Nascent chains[b]	Ribosomes[c]
Co-translationally targeted[d]				
Endoplasmic reticulum	113	446	7%	19%
Cell membrane	56	494	6%	15%
Lysosome/Vacuole	30	482	2%	7%
Post-translationally targeted[e]				
Extracellular	13	246	79%	44%
Cell membrane	9	267	2%	3%
Endoplasmic reticulum	7	453	0%	1%

[a] Calculated as the average number of amino acids

[b] Calculated as percentage of total cTPM for all proteins predicted to enter the ER

[c] Calculated as percentage of total cRPM for all proteins predicted to enter the ER

[d] Greater than 2-fold membrane enrichment and not mitochondrial

[e] Lesser than 2-fold membrane enrichment, not mitochondrial, and contain a signal sequence

2.3.2 Comparing the translational landscape of *K. phaffii* and *S. cerevisiae*

Of the 5,329 *K. phaffii* genes annotated here, 73 % have a homolog in *S. cerevisiae*. Unlike *K. phaffii*, *S. cerevisiae* is thought to have undergone a whole-genome duplication, and so many *S. cerevisiae* genes have paralogs.[111] The influence of paralogy is evident in how these two species allocate translational throughput. We calculated cTPM and cRPM in *S. cerevisiae* using prior data acquired under similar growth conditions.[21,35] The overall distribution of cTPM by ontological category is similar between species (*Figure 2.5*). Under the conditions tested here (glucose-containing rich media), *TEF1*, encoding translational elongation factor 1 alpha, is the most translated protein in *K. phaffii*. The *TEF1* promoter is used to drive constitutive expression in *K. phaffii*,[112] and our results suggest that the native *TEF1* ORF is translated more than the ORFs linked to other promoters used for expression in glucose, such as *GAP* (here, *TDH3*) and *PGK1*.[12] *S. cerevisiae* generates a similar amount

Figure 2.5: **Comparison of metabolic burden for *K. phaffii* and *S. cerevisiae* a.** Total nascent chains for *K. phaffii*. **b.** Total nascent chains for *S. cerevisiae*.

of nascent chains to the same function, but it does so using a combination of its paralogous genes *TEF1* and *TEF2*. Unsurprisingly, Crabtree-positive *S. cerevisiae* generates three times more polypeptides involved in glycolysis and fermentation than *K. phaffii* (e.g., *ENO1/2*, *GMP1*, *FBA1*, *TDH2/3*, *TPI1*, *PGK1*, *PDC1*, *ADH1*).

Indeed, these two species also show divergence in energy production with regards to co-translational mitochondrial import (*Figure 2.6*). Our subcellular frac-

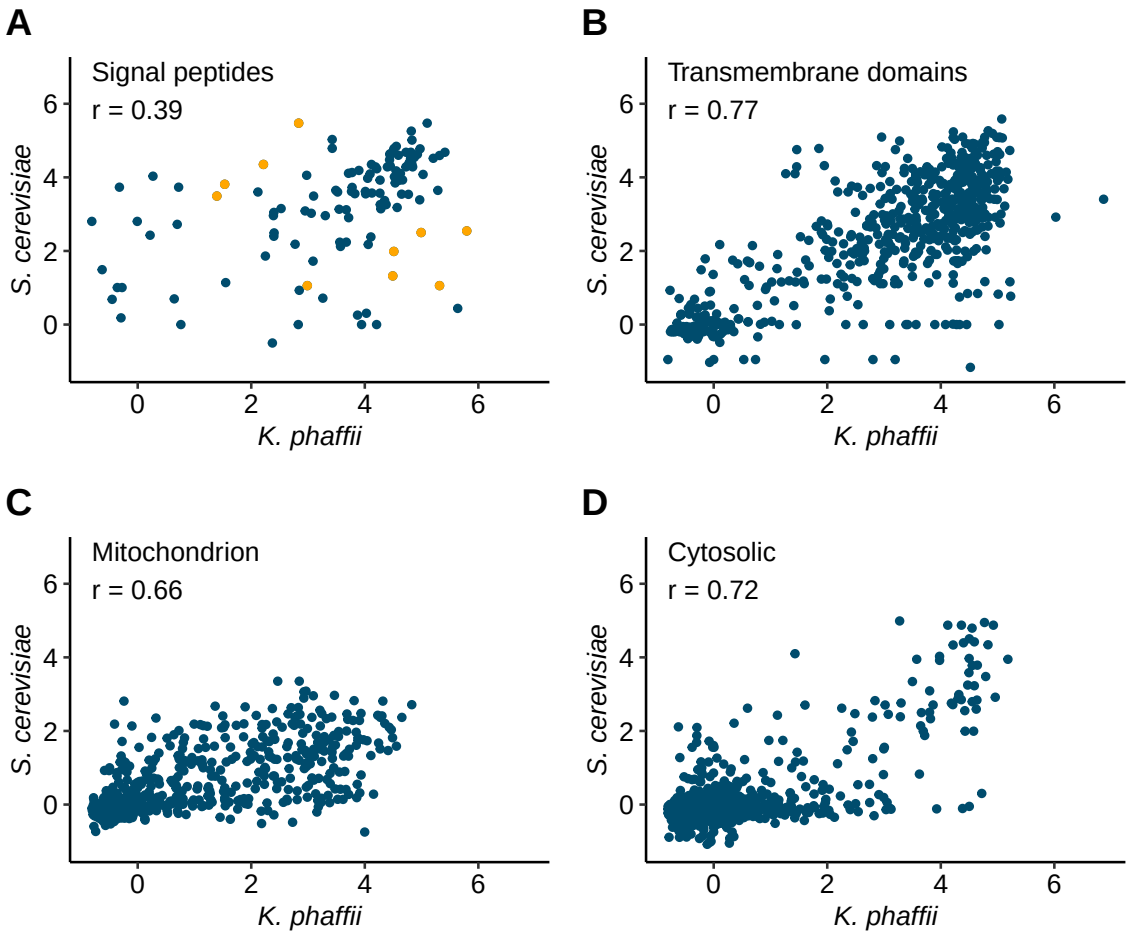

Figure 2.6: **Correlation of membrane enrichment scores between *K. phaffii* and *S. cerevisiae*** *K. phaffii* used membrane enrichment scores from a biological replicate not treated with CHX and *S. cerevisiae* used membrane enrichment scores from a biological replicate treated with CHX. ***a.*** Enrichment scores restricted to non-mitochondrial signal sequence bearing proteins. Contrast dots represent genes found in *Table 2.3*. ***b.*** Enrichment scores restricted to non-mitochondrial transmembrane proteins. ***c.*** Enrichment scores restricted to mitochondrial proteins. ***d.*** Enrichment scores restricted to cytosolic proteins.

tionation assay recovers all membrane-bound ribosomes, including those attached to the mitochondria. A greater number of nuclear-encoded mitochondrial proteins undergo membrane-localize translation in *K. phaffii*. Recovery of membrane associated mRNA strongly depends on active translation.[21] Therefore, less active

translation of mitochondrially destined proteins may become reflected in lower membrane-enrichment scores.

We next asked whether ER translocation pathways are conserved between the two species. Between homologs, membrane enrichment scores correlated with a Pearson's r of 0.85 compared to a Pearson's r of 0.99 between *K. phaffii* replicates (*Figure 2.7*). Genes encoding transmembrane proteins or cytosolic proteins

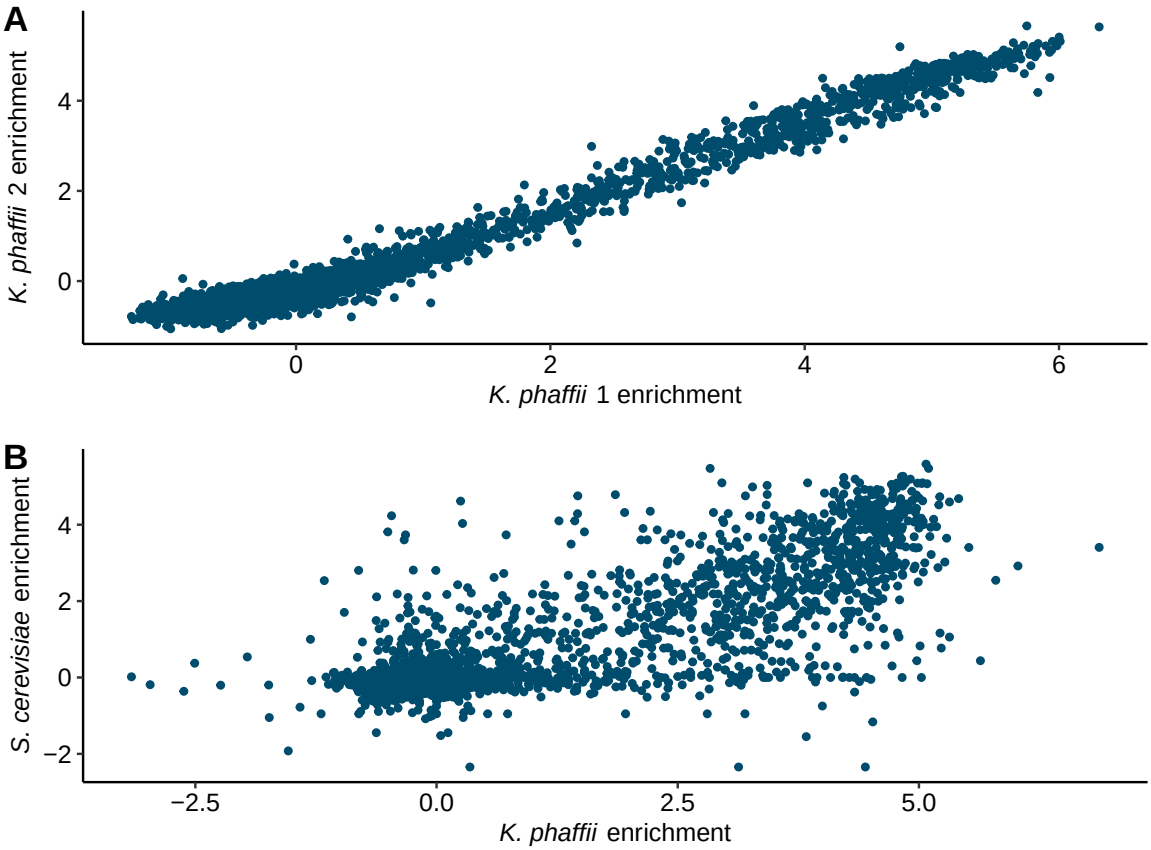

Figure 2.7: **Comparison of membrane enrichment between *K. phaffii* and *S. cerevisiae*** ***a*** Comparison of membrane enrichment between *K. phaffii* replicates. ***b*** Comparison of membrane enrichment between *K. phaffii* and *S. cerevisiae*.

which lack ER or mitochondrial targeting sequences had the highest correlation. Signal-sequence bearing proteins, including GPI-anchored proteins, however, had lower correlation (*Figure 2.7a*). There were several genes which only showed co-translational membrane enrichment in one species, and in some cases this was

Table 2.3: **Membrane enrichment for secreted, lumenal and GPI-anchored proteins in *K. phaffii* and *S. cerevisiae***

Gene	Description	*K. phaffii*	*S. cerevisiae*
Increased enrichment			
FLO9	Lectin-like protein, flocculin (isoform 2)	5.32	1.06
ZPS1	Putative GPI-anchored protein	5.80	2.54
SGA1	Sporulation-specific glucoamylase	4.49	1.32
BIG1	Cell wall beta-1,6-glucan level regulator	4.51	1.99
GDA1	Guanosine-diphosphatase	4.99	2.50
FLO9	Lectin-like protein, flocculin (isoform 1)	2.99	1.06
Decreased enrichment			
YKL077W	Uncharacterized protein	1.39	3.49
PDI1	Protein disulfide isomerase	2.21	4.35
MNL1	Uncharacterized protein	1.53	3.81
KRE5	Beta-1,6-glucan biosynthesis protein (isoform 2)	2.84	5.47

due to lost of a signal peptide in one of the homologs. The ten genes that showed the greatest difference in magnitude, while still showing evidence for membrane enrichment in both species, are reported in *Table 2.3*. Notably, this list includes *PDI1*, encoding an ER lumenal protein-disulfide isomerase that is essential for ER homeostasis. Mitochondrially localized proteins have greater membrane enrichment in *K. phaffii*, which may be related to the greater use of aerobic respiration compared to *S. cerevisiae* (*Figure 2.7c*).

Finally, we explored the relationship between the burden imposed by production of polypeptide chains (cTPM), ribosome demand (cRPM) and translocation pathway (membrane enrichment score) for ER destined proteins within the two species (*Figure 2.8*). In *S. cerevisiae*, most of these chains originate from a single gene, *CCW12*, while in *K. phaffii*, there are a wider variety of genes, with *SCV12161.1* being the most dominant. Strikingly, post-translational targeting is used for about two-thirds of lumenal, secreted or GPI-anchored nascent chains in both species. *K. phaffii*, however, is distinguished by at least one major cell wall protein, Pst1p, which enters the ER co-translationally. In both species, Pma1p is the dominant

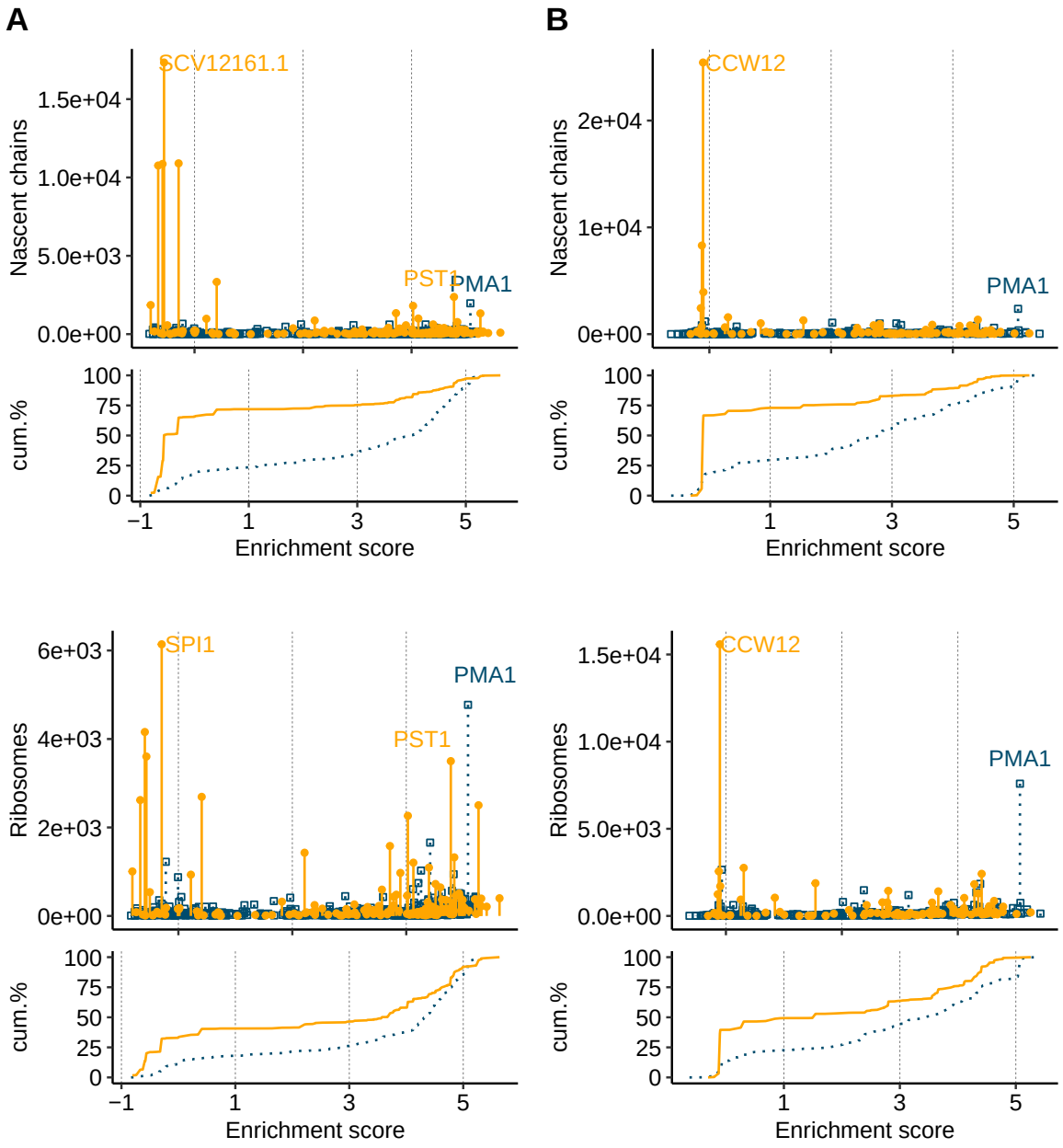

Figure 2.8: **Demands imposed on secretion pathway** Black dotted lines represent membrane proteins and blue solid lines represent secreted, lumenal or GPI-anchored proteins. Membrane proteins were non-mitochondrial proteins that contained greater than or equal to two transmembrane domains or had one transmembrane domain but do not have a predicted N-terminal signal sequence. Secreted, lumenal or GPI-anchored proteins proteins were non-mitochondrial proteins containing a predicted N-terminal signal sequence and less than or equal to one transmembrane domain. ***a*** Demands in *K. phaffii*. ***b*** Demands in *S. cerevisiae*.

membrane protein passing into the ER. In terms of ribosome sequestration, the trend reverses; co-translational translocation is responsible for sequestering two thirds of ribosomes used to produce secreted or GPI-anchored proteins. While *PST1* yields slightly more nascent chains than *PMA1*, *PMA1* is more than twice as long as *PST1* and sequesters 1.36 times more ribosomes. Thus, *PMA1* represents a significant burden to the secretory systems of both *S. cerevisiae* and *K. phaffii* as it is predicted to sequester more ribosomes, co-translational translocons, and lumenal chaperones to synthesize and transport nascent chains into the ER.

2.4 Discussion

The yields of engineered, recombinant proteins are restricted by bottlenecks in biogenesis.[2] Certain bottlenecks are metabolic, including insufficient ATP or other high-energy compounds, nucleotides for mRNA synthesis, amino acids, carbohydrates for glycosylation, and reducing equivalents. A promising systems-level approach to remove bottlenecks is to identify and delete host proteins with the greatest demand for metabolic resources. Indeed, the Lewis lab has elegantly demonstrated in CHO cells that deleting expensive proteins (in terms of ATP equivalents) increases the yield of heterologous secreted proteins.[26,107,108] Similar modeling of metabolic demand has been performed by the Nielsen lab for the secretome of *S. cerevisiae*.[106] Other bottlenecks are due to insufficient cellular protein biosynthetic machinery, such as polymerases, ribosomes, translocons, and molecular chaperones. Focusing on metabolic demand will likely relieve pressure on machinery with tightly coupled–and therefore accurately predicted–energetic requirements (e.g., cycles of translation elongation by the ribosome). However, it only approximates demand for chaperones and translocons, which gate entry into the ER. Compared to tightly

coupled complexes, chaperones and translocons are ambiguous in their energetic demand. Chaperones perform cycles of binding and rebinding that depend on the folding pathways of client proteins.[113] Translocation into the ER is driven by ATP-hydrolyzing chaperones, translation elongation, or a combination of the two in a client dependent manner.[114,115] Engineering of the early secretory pathway, such as the optimization of signal sequences for protein targeting[27] and reducing the effect of the ERAD system,[20] provides varying degrees of success. These approaches are contingent on the complexity of the protein product and must be empirically optimized.[28,29] Our data and analysis may augment these efforts by accounting for capacity of translation, co- and post-translational translocation.

Despite the ability of Ribo-seq to accurately quantify gene expression, our study has several caveats that limit interpretation. First, we have only considered yeast undergoing log phase growth in liter scale, aerated shaking cultures using rich media. This design enabled comparison to several published data sets using *S. cerevisiae* that were collected under identical conditions.[21,35] We chose strain GS115, a commonly used commercially available strain that is a histidine auxotroph (*HIS4*). Even under rich media with abundant extracellular histidine, this auxotrophy may influence gene expression compared to strains which supply His4p. Future work involves quantifying demands at industrial scale in stirred bioreactors under induction of a heterologous protein. Second, we assume that elongation rates are relatively constant across genes. However, if the elongation rate is altered for a transcript, it may result in greater or fewer ribosome protected reads. We argue that on the whole, our assumption is valid, given that Ribo-seq accurately predicts mature protein stoichiometry.[35,116] Third, Ribo-seq does not account for protein degradation; indeed, some proteins are co-translationally ubiquitinated.[117] Our results should therefore not be interpreted as revealing steady-state protein levels in

K. phaffii. However, our goal was to quantify the costs of protein synthesis, and so we argue that Ribo-seq is a more appropriate tool than mass spectrometry. Despite these limitations, our approach allowed us to interrogate protein translocation into the ER.

Most secreted proteins, including high-value targets like antibodies, will enter the ER via a sec translocon.[3] The translocon subunits Sec62p, Sec63p, Sec66p and Sec72p are required for the translocation of certain proteins, particularly those with shorter or less hydrophobic signal peptides.[23,95,99] Molecular chaperones are also implicated in protein translocation, through binding of proteins in the cytoplasm (Ssa1p)[94] or the ER lumen (Kar2p).[114] However, many gene products are able to associate with more than one class of translocon.[23,99] In addition, while recent structural work suggests that the heptameric Sec61 complex cannot directly bind a ribosome,[25,118] there is a preponderance of evidence demonstrating that the proteins dependent on this complex are translated at the ER membrane.[21,23,98,119,120] Further, even if a protein does not strictly require particular machinery, like SRP, it may nonetheless sequester it in vivo, reducing availability for proteins that do require these factors.[21,119] Because of these complexities, it is unsurprising that it has remained difficult to precisely tune a translocon for a specific engineered protein. Rather, optimization will likely require understanding the needs of the target, what the target will sequester, and how this will relate to the balance of resources in the host.

Our calculations for nascent chains produced, ribosomes used, and predicted translocation pathways suggest that each gene presents a unique combination of challenges to the cellular biosynthetic capacity. For instance, long, co-translationally translocated proteins will impart little demand on cytoplasmic chaperones, but will sequester ribosomes, translocons, and lumenal chaperones for extended periods

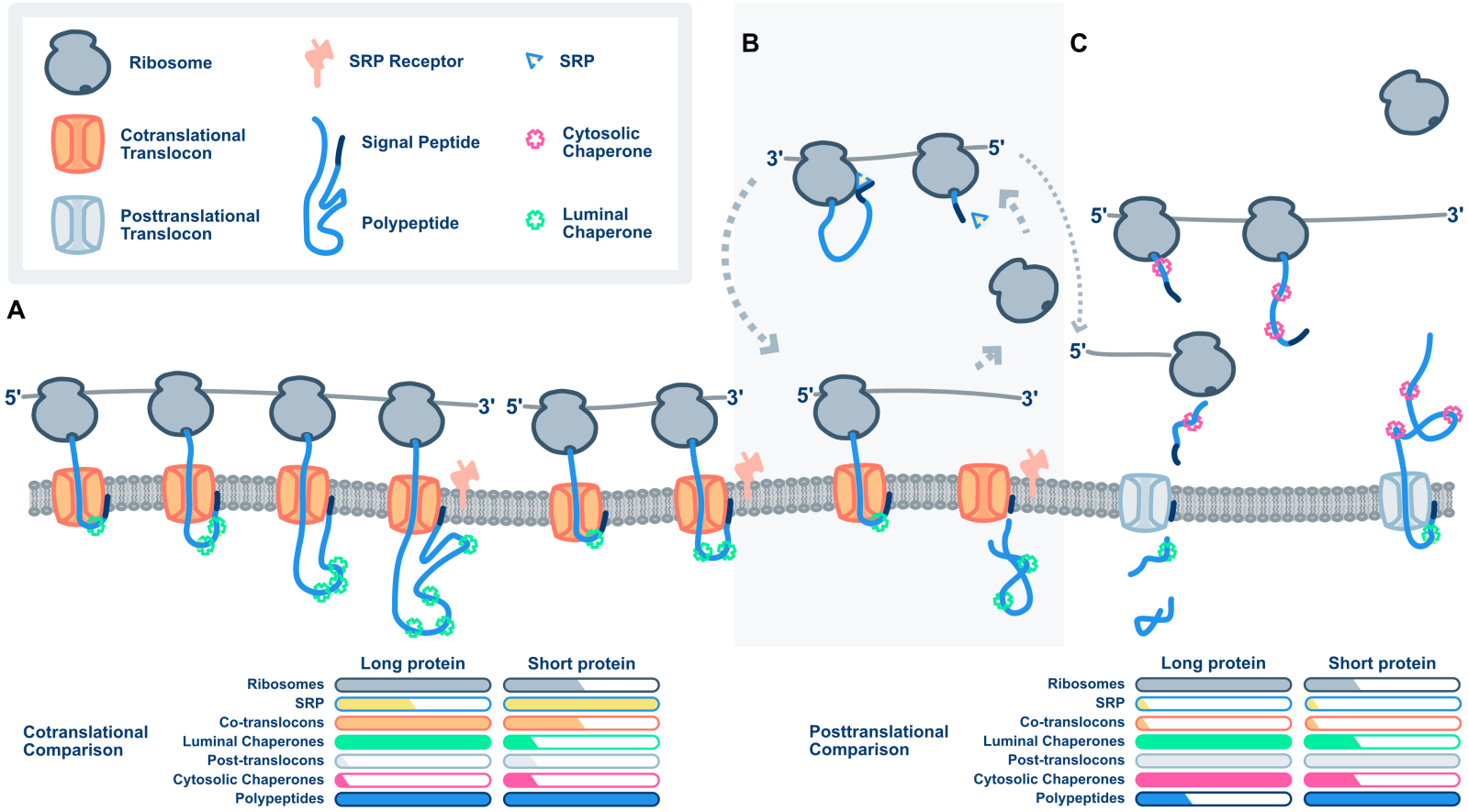

Figure 2.9: **Demands imposed by different translocation pathways** *a* Co-translational translocation of a long and short protein. Shorter proteins translate faster and allow resources to be released and recycled. This allows shorter genes to use less resources to produce the same amount of proteins as a longer gene in a given time. *b* Alternative pathway for co-translationally destined short proteins. Genes that are too short to target to the ER in time may also use the post-translational pathway for ER translocation. *c* Post-translational translocation of a long and short protein. Shorter genes produce more nascent chains than a longer gene given the same amount of time and resources.

of time (*Figure 2.9a*). However, because of sustained translation on the surface of the ER, fewer instances of SRP targeting are required. A shorter co-translational protein will require fewer ribosomes, translocons, and lumenal chaperones to produce the same number of polypeptide chains. However, if the gene is short enough to fail to sustain translation at the membrane (*Figures 2.3, 2.9*), then it may require multiple rounds of SRP targeting to get there. If sufficient nascent chains are exposed to the cytosol, the gene may also require cytosolic chaperones. If translation terminates prior to membrane attachment, then post-translational translocons may be needed as well. Long, post-translationally translocated proteins will also sequester ribosomes, but will require both lumenal and cytosolic chaperones (*Figure 2.9c*). There are few genes in *K. phaffii* in this category (*Figure 2.3*). Finally, short, post-translationally translocated proteins will sequester few ribosomes, no co-translational translocons, and some cytosolic and lumenal chaperones. Our experimental approach cannot measure transit time through post-translational translocons; we speculate that it will be correlated to polypeptide length.

Some resources used in biogenesis of ER proteins are dependent on chain number, rather than elongation time. For instance, GPI-anchored proteins each receive a single lipid anchor,[121] retrograde transport is mediated by the K/HDEL recognition,[122] and protein sorting in the secretory pathway involves interactions between cargo and receptors, such as Sec24p.[123] In optimizing these systems, cTPM may be the appropriate metric to consider, and strain engineering efforts could focus on deleting or down-regulating highly expressed host proteins. In yeasts, GPI-anchored cell wall proteins present the greatest burden by cTPM. Other aspects are dependent on total polypeptide length, such as the potential ratcheting mechanism provided by Kar2p during translocation.[114] Although not considered here, cTPM scaled by protein length may be the appropriate metric used in engineering. A

third aspect is the availability of resources such as ribosomes or translocons, which are sequestered while in operation. cRPM is an appropriate metric to understand ribosome sequestration. For co-translational translocation, we propose that cRPM could be used as a proxy, as one ribosome binds one translocon during import. In *S. cerevisiae* and *K. phaffii*, expression of *PMA1* appears to be a major ribosome sink, and therefore also a translocon sink. In *K. phaffii*, *PST1* is a second major sink for ribosomes and translocons.

Although fungi are genetically and physiologically diverse, most mechanistic knowledge about secretion is derived from studies in *S. cerevisiae*.[3] Based on a recent molecular dating using 332 genomes,[124] *K. phaffii* and *S. cerevisiae* diverged roughly 230 million years ago, whereas the *S. cerevisiae* whole-genome duplication occurred roughly 90 million years ago. Thus, sequence variation is found in nearly all of the proteins conserved in the two species, and due to the paralogy in *S. cerevisiae*, additional differences exist in the regulation of gene expression. Our comparison of *K. phaffii* and *S. cerevisiae* suggests that the path a conserved protein takes to the ER is not necessarily the same between species, even for essential genes critical to health of the secretory pathway, like *PDI1*. However, we find that even though the number and diversity of genes differ between the species, categorically there is conservation in the biosynthetic demand. For instance, our data suggest that *K. phaffii* can provide more nuanced engineering of the cell wall, as it is composed by a greater number of genes. Optimizing fungal species separately may increase protein secretion yields in ways not predicted through analysis of model organisms alone. These results call for a more thorough understanding of industrially used fungal secretion systems for rationally engineering cellular factories during bioproduction.

2.5 Conclusions

Protein biogenesis is a complex phenomena that not only requires raw materials (energy and amino acids), but also access to specialized cellular machinery. Our analysis in *K. phaffii* reveals several principles about these pathways that will be useful in strain engineering. First, we find that a small number of host genes are responsible for most of the protein entering the secretory pathway. Second, GPI-anchored protein components of the cell wall represent the greatest number of nascent chains within the secretory pathway. Third, co-translational translocation pathways must accommodate a wider set of proteins than post-translational pathways. Fourth, orthologs may enter the endoplasmic reticulum through different translocation pathways. Finally, despite differences in the number of genes associated with biological function, the amount of nascent chains entering the ER are similar between *K. phaffii* and *S. cerevisiae*.

2.6 Availability of data and materials

The datasets generated and analyzed during the current study are available as NCBI Bioproject PRJNA669501.

Chapter 3

Identification of targets for rational strain engineering in *Komagataella phaffii* using ribosome profiling

Background: Biogenesis of heterologous proteins requires the harmonization of biological machinery to synthesize, fold, traffic, and secrete proteins. However, demands imposed by the host proteome present significant bottlenecks for heterologous proteins to undergo these processes. To understand these demands, we quantified the translatome of *Komagataella phaffii* under heterologous conditions involving growth on glycerol and methanol medias.

Results: We optimized ribosome profiling for studying translation in heterologous conditions using combinatorial ribosomal RNA depletion strategies and develop an approach for differential expression analyses that relies on minimal numbers of replicates. In using ribosome profiling, we quantified global and early secretory demands before and after methanol induction in *Komagataella phaffii* GS115 Mut^S *ALB* and GS115 Mut^+ strains. For global demands, both strains show distinct

patterns of translation initiation, elongation, and gene expression before and after methanol induction that were indicative of oxidative stress responses. Surprisingly, GS115 Mut^S *ALB* shows lesser expression of genes involved in the unfolded protein response and endoplasmic reticulum associated degradation than GS115 Mut$^+$. Protein components of the cell wall represent the greatest number of nascent chains entering the early secretory pathway for both strains. However, cell wall components were significantly more expressed in GS115 Mut^S *ALB* than GS115 Mut$^+$. Additionally, the most highly expressed cell-wall components before induction were distinct from those after induction.

Conclusion: Ribosome profiling requires condition specific optimizations to capture translation most accurately. Oxidative stress may limit bioproduction of albumin-like proteins more than heterologous folding stress. Bottlenecks involved in methanol induced protein production may be alleviated by optimizing media methanol concentrations, over expressing genes involved in combating oxidative stress, and deleting host cell proteins that sequester the most biosynthetic resources in the early secretory system.

3.1 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[5] and have the capacity to treat a variety of ailments including but not limited to central nervous system disorders,[125] inflammation,[126] hypercholesterolemia,[127] and infectious diseases.[128] The ability to produce these proteins remains difficult and is viable to the biogenetic capacity of production

strains. Fungi are used as production strains because of their ability to grow rapidly to high densities in inexpensive media, are easy to genetically manipulate, and have the ability to post-translationally modify proteins.[2,4]

Komagataella phaffii stands out among the fungal kingdom for its ability to metabolize methanol as its primary carbon source using the alcohol oxidase (AOX) and for the limited host proteins it naturally secretes into its media.[11,105,129] Production of AOX is constitutively expressed by *AOX1* in the absence of glucose and the presence of methanol.[130] Industrial bioproduction in *K. phaffii* typically involves growing cells in glycerol-based media before transferring them to methanol-based media for heterologous induction.[47,131,132] This process utilizes the *AOX1* promoter to precisely regulate the expression of heterologous proteins like human serum albumin (HSA).[13] HSA is moderately sized (~67 kDa) protein with semi-complex folding requirements and is minimally glycosylated. As an industrially relevant recombinant protein, HSA is a major protein component of human plasma and is produced as a serum replacement product to maintain colloid osmotic pressure within blood vessels. The cost and sophistication of producing HSA is reduced if the host cell secretes it into the growth media as this simplifies downstream purification.[2,19,133,134] However, protein secretion is complex and viable to multiple bottlenecks.

The first major bottleneck, protein trafficking through the endoplasmic reticulum (ER), is complex and is the rate limiting step in protein production.[20] Heterologous trafficking is contingent on the recognition and binding of N-terminus hydrophobic motifs, signal sequences, by a signal recognition particle (SRP).[135] 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.[22] In *S. cerevisiae*, proteins translocate co-translationally using

the hexameric and heterotrimeric Sec-translocon. Secreted proteins that fail to translocate across the ER do not have access to ER-resident chaperones and do not fold correctly. Protein folding is an ATP driven process that includes many ER-resident proteins such as Kar2p, Scj1p, Pdi1p, Ero1p, and Jem1p. Access to protein folding chaperones in *K. phaffii* is made more difficult as previous studies show that an equal amount nascent polypeptides translocate across the ER co-translationally and post-translationally.[136] 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).[137]

As a strategy to improve secretion in *K. phaffii*, we propose the use of next generation sequencing to study the translatome under heterologous conditions for rational strain engineering purposes. We provide analyses of methanol metabolism in *K. phaffii* for insights into process optimization. As well, we show which host cell proteins sequester the most biogenetic machinery in the early secretory pathway during heterologous expression. These insights are accomplished using ribosome profiling (Ribo-seq) and ER trafficking predictions to produce data sets that reflect prototypical variations in the translatome under heterologous conditions in wild-type and HSA expressing strains. We provide a model for metabolic and secretory demands by surveying proteins expressed globally and proteins that are predicted to enter the ER co-translationally and post-translationally. Little is known of how host protein synthesis changes under heterologous conditions and lesser is understood how the cell manages resources to traffic and translocate engineered proteins. Our experiments reveal novel insights into these conditions and may allow for

a rational approach to widen secretion bottlenecks by providing new targets for modification that would not have otherwise been predicted.

3.2 Materials and Methods

3.2.1 Strains and culture conditions

Assays were performed using GS115 Mut$^+$ and GS115 Mut^S *ALB*.[13]

For each biological replicate, $200\,\mathrm{mL}$ liquid cultures of BMGY ($1\,\%$ yeast extract, $2\,\%$ peptone, $100\,\mathrm{mm}$ potassium phosphate pH 6.0, $1.34\,\%$ YNB, and $1\,\%$ glycerol) were grown to an $\mathrm{OD}_{600\,\mathrm{nm}}$ of 5 at $30\,^{\circ}\mathrm{C}$ with shaking in baffled $2\,\mathrm{L}$ flasks. Of this culture, $100\,\mathrm{mL}$ were harvested by vacuum filtration through a $0.8\,\mu\mathrm{m}$ filter. Immediately after filtering, cells were scraped off the filter using a chilled scoopula and submerged in a $50\,\mathrm{mL}$ conical tube containing liquid nitrogen. The remaining liquid cultures were split into two $50\,\mathrm{ml}$ conical tubes and were pelleted via centrifugation. Supernatant was removed from each $50\,\mathrm{ml}$ conical tube. The cell pellet of one $50\,\mathrm{ml}$ conical tube was gently resuspended with $40\,\mathrm{mL}$ BMMY without methanol ($1\,\%$ yeast extract, $2\,\%$ peptone, $1.34\,\%$ YNB, and $100\,\mathrm{mm}$ potassium phosphate pH 6.0). Resuspended culture was used to resuspend the cell pellet in the second $50\,\mathrm{ml}$ conical tube. Resuspended cultures were equally divided into two $280\,\mathrm{mL}$ cultures of BMMY without methanol in $2\,\mathrm{L}$ baffled flasks for a final volume of $300\,\mathrm{mL}$ for each sample. Methanol was added at $0.5\,\%$ to each of the baffled flasks for *AOX1* induction. Flasks were allowed to shake at $30\,^{\circ}\mathrm{C}$ and were collected in the manner described above three and twenty-four hours after methanol induction (*Figure 3.1*). Lysis buffers ($50\,\mathrm{mM}$ MOPS, $25\,\mathrm{mM}$ KOH, $100\,\mathrm{mM}$ KOAc, $2\,\mathrm{mM}$ MgOAc, $1\,\mathrm{mM}$ DTT, and $1\,\%$ Triton X-100) for each sample were frozen by adding $2\,\mathrm{mL}$ dropwise

to a $50\,\mathrm{mL}$ conical tube containing liquid nitrogen. For each sample, frozen cells were mixed with $2\,\mathrm{mL}$ frozen lysis buffer. Cell fractions were pulverized for $2\,\mathrm{min}$ in a $50\,\mathrm{mL}$ ball mill chamber with a single $2\,\mathrm{cm}$ steel ball (Retsch) and collected in $50\,\mathrm{mL}$ conical tubes. After thawing, lysates were centrifuged at $18\,000\,\mathrm{g}$ for $10\,\mathrm{min}$. Supernatants were transferred to $1.5\,\mathrm{mL}$ conical tube and were further clarified by centrifugation at $23\,000\,\mathrm{g}$ for $20\,\mathrm{min}$.

3.2.2 Ribo-seq

Lysed samples were nuclease digested using $40\,\mathrm{U}$ of Ambion RNase A for $1\,\mathrm{h}$ at room temperature. Digested samples were layered on a $10\,\%$ to $50\,\%$ sucrose gradient prepared in $50\,\mathrm{mM}$ Tris pH 7.5, $200\,\mathrm{mM}$ NaCl, and $2\,\mathrm{mM}$ MgOAc case using a Gradient Master (Biocomp). Gradients were centrifuged at $39\,000$ RPM for $2.5\,\mathrm{h}$ 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 $18\,\mathrm{nt}$ to $34\,\mathrm{nt}$ were resolved and excised using $15\,\%$ polyacrylamide TBE-urea gel. RNA was collected from excised gel fragments using RNA gel extraction buffer ($300\,\mathrm{mM}$ NaOAc, $1\,\mathrm{mM}$ EDTA, and $0.25\,\%$ SDS), precipitated, and resuspended in water containing $20\,\mathrm{U\,mL^{-1}}$ SUPERase · In.

Purified fragments were then dephosphorylated by incubating $2\,\mu\mathrm{L}$ $1\,\mathrm{M}$ RNA sample with $2\,\mu\mathrm{L}$ RNase free water, $0.5\,\mu\mathrm{L}$ SUPERase · In RNase Inhibitor, $0.5\,\mu\mathrm{L}$ T4 Polynucleotide Reaction Buffer (PNK), and $0.5\,\mu\mathrm{L}$ T4 Polynucleotide Kinase at $37\,^{\circ}\mathrm{C}$ for $1\,\mathrm{h}$. Dephosphorylated samples were linker ligated with adapter sequences by incubating with $3.5\,\mu\mathrm{L}$ $50\,\%$ PEG-8000, $0.5\,\mu\mathrm{L}$ 10X T4 RNA Ligase Reaction Buffer, $0.5\,\mu\mathrm{L}$ $10\,\mu\mathrm{M}$ adenylated linkers and $0.5\,\mu\mathrm{L}$ T4 Rnl2(tr)k277Q at $30\,^{\circ}\mathrm{C}$ for

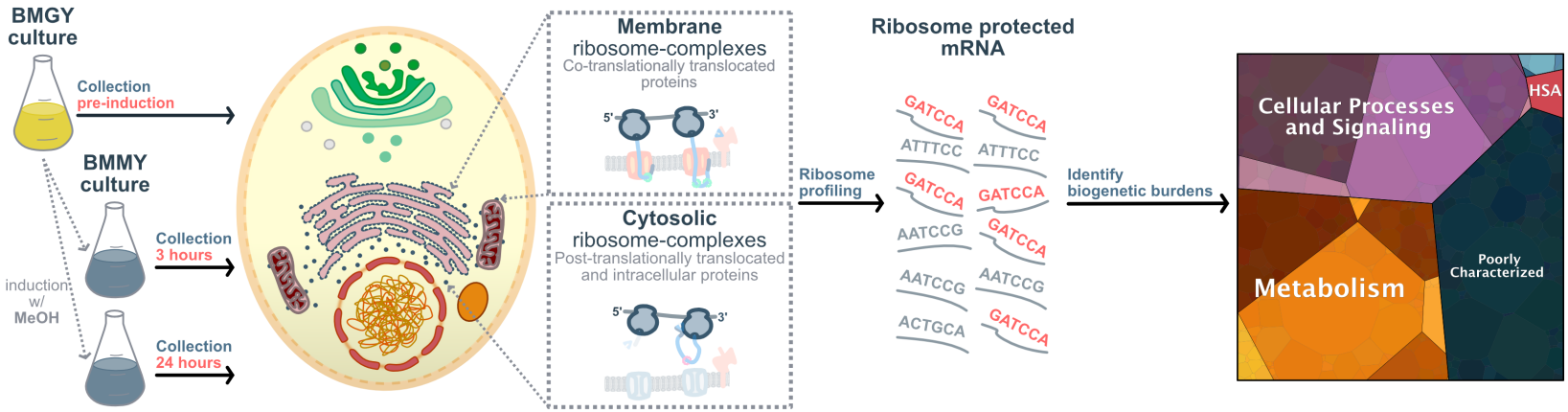

Figure 3.1: **Overview of heterologous expression and Ribo-seq** Starter cultures were grown in buffered glycerol media (BMGY). Approximately $\frac{1}{3}$ of starter culture was collected via vacuum filtration and flash freezing. The remaining $\frac{2}{3}$ of starter culture was decanted before heterologous induction using buffered methanol media (BMMY). Induced culture was split into equal volumes before each subculture was collected 3 and 24 hours afterwards. Membrane-associated and cytosolic ribosomes were isolated from cell lysates. Membrane-associated ribosomes include co-translationally translocated proteins on the mitochondria and endoplasmic reticulum (ER). Cytosolic ribosomes include intracellular proteins and proteins that post-translationally translocate into the ER. mRNA footprint are isolated from ribosomes before Illumina sequencing. Ribo-seq libraries reveal host protein synthesis under heterologous conditions.

4 h. Linker-ligated samples were concentrated via isopropranol 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 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 method previous 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 sample at 80 °C for 10 min on thermal block.

Circularized samples were rRNA depleted, again, using probe-directed degra-

dation via double stranded nuclease (DSN).[138,139] Depletion probes were designed using rRNA aligned Ribo-seq reads collected from GS115 Mut^S *ALB* cultured in BMGY before methanol induction (*Table 3.1*). Circularized samples ($10\,\mu L$) were incubated with $4\,\mu L$ 4x hybridization buffer, $1\,\mu L$ 4x depletion probes at $200\,\mu M$, and $1\,\mu L$ 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 $2\,\mu L$ 10x DSN master buffer, $1\,\mu L$ DSN storage buffer, and $1\,\mu L$ DSN before incubation at $65\,°C$ for $25\,min$. Reaction was stopped by adding $20\,\mu L$ $10\,mM$ 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 $50\,\mu L$ reaction mixture ($10\,\mu L$ Q5 Reaction Buffer, $1\,\mu L$ $10\,mM$ dNTPs, $2.5\,\mu L$ $10\,\mu M$ forward primer, $4\,\mu L$ circularized DNA sample, $0.5\,\mu L$ Q5 High Fidelity DNA Polymerase and $29.5\,\mu L$ RNAse free water) divided into 5 x $10\,\mu L$ 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.

3.2.3 Mapping of ribosome-protected reads to codons

Sequenced reads were trimmed and demultiplexed in an error-tolerant way using Cutadapt.[65,140] Reads were computationally rRNA subtracted by aligning them to *Komagataella pastoris* GS115 genomic rRNA using HISAT2.[56,66] Subtracted reads were

mapped to the genome for *Komagataella pastoris* GS115[46] 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.[67,141] Genomic alignments were loaded into R using the GenomicAlignments package from Bioconductor.[68] 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[136] and the RiboProfiling package in Bioconductor.[69] Genomic alignment objects were used with p-site offsets to generate reads per codon per gene (RPCPG) data tables.

3.2.4 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 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 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.[45]

3.2.5 Normalization and differential expression analysis

Read counts were normalized at the codon level using a metagene correction strategy previously discussed in[136] with some modification. Reads for the first 500 codon positions at the 5'end of all transcripts were 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 of Ingolia et al..[79] 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 expression 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.

3.2.6 Classification of ORFs

Open reading frames for each genes were characterized using various prediction softwares: clusters of orthologous groups were predicted using EggNOG 4.5,[73] subcellular localization was predicted using DeepLoc,[75] signal sequences were predicted using SignalP 5.0,[76] transmembrane domains were predicted using TOPCONS,[77] and GPI-anchors were predicted using predGPI.[78] 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.[136] These data sets revealed expression 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 more broad 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 lesser 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.

3.3 Results

3.3.1 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, three hours after, and twenty-four hours after methanol induction. Ribo-seq utilizes a non-specific nuclease to degrade nucleic acids, including mRNA, that are not covered and therefor protected by ribosomes. In order to sequence ribosome protected mRNA fragments and reveal translational dynamics, ribosome derived RNA 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[136] 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[79] and revealed that a subset of rRNA represented the majority of rRNA contamination (*Table 3.1*). From the pre-induction sample, complimentary oligos of this subset were used for probe-directed degradation using double stranded nuclease (DSN). Using probe-directed DSN treatment, rRNA contamination was

Table 3.1: **Oligos designed for probe-directed degradation**

Probe sequence[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] Complentary oligonucleotides to Ribo-seq sequences mapped most highly to GS115 rRNA
[b] Ribo-seq reads aligned to GS115 rRNA

reduced from $88\,\%$ to $10\,\%$ in the pre-induction sample, $87\,\%$ to $20\,\%$ in the 3 hour post-induction sample, and $93\,\%$ to $62\,\%$ in the 24 hour post-induction sample. Before induction 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[79] (*Figure 3.2*).

Our data revealed genome wide coverage of expression as up to $96\,\%$ of *K. phaffii*'s $5,330$ annotated protein-encoding genes were detected. Before making intra- and inter-sample comparisons of expression levels, we first sought to normalize reads (*Figure 3.3*). First, footprint sized fragments were used to generate compu-

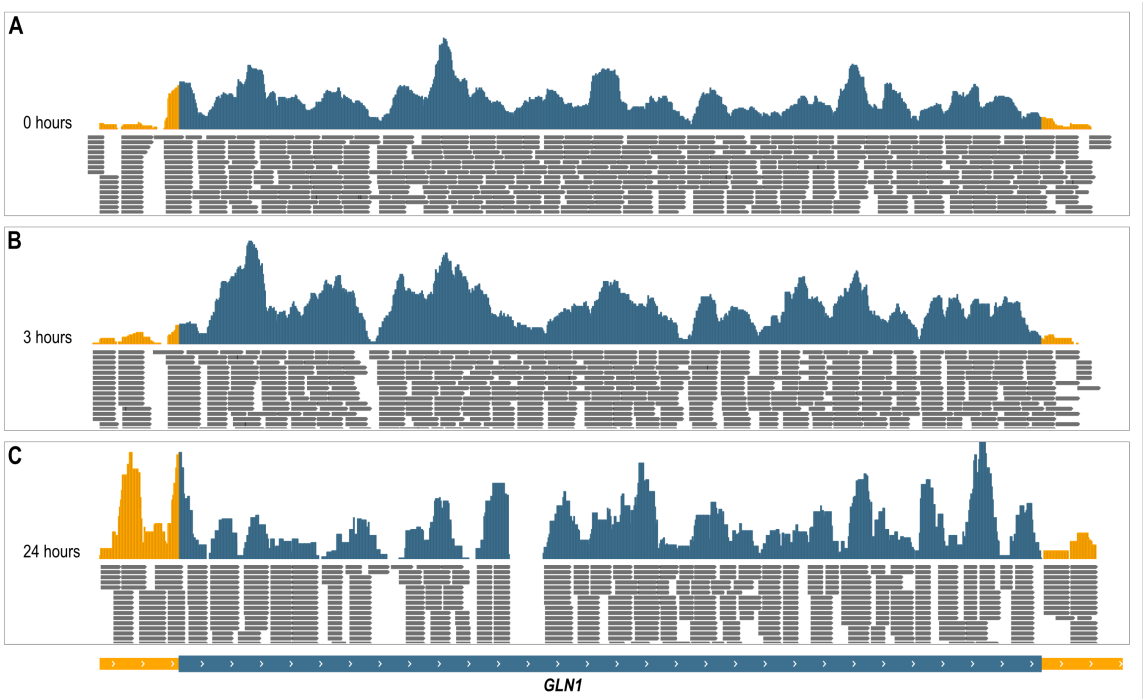

Figure 3.2: **Ribosome abundance on transcripts under heterologous conditions** Images are modified screen captures from Integrated Genome Viewer (MIT). All registers represent transcript reads from GS115 Mut^S *ALB* cultures collected at different moments. The bottom band shows predicted transcript boundaries for *GLN1*. The thick blue band shows the open reading frame (ORF) while the thinner yellow band shows the untranslated regions (UTRs). ***a*** In the pre-induction sample cultured BMGY media, the majority of reads map to ORFs. ***b*** Three hours after induction in BMMY media, the majority of reads still map to ORFs. ***c*** After 24 hours of heterologous expression, a much higher proportion of reads map to the 5'UTR.

tational 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 expression 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

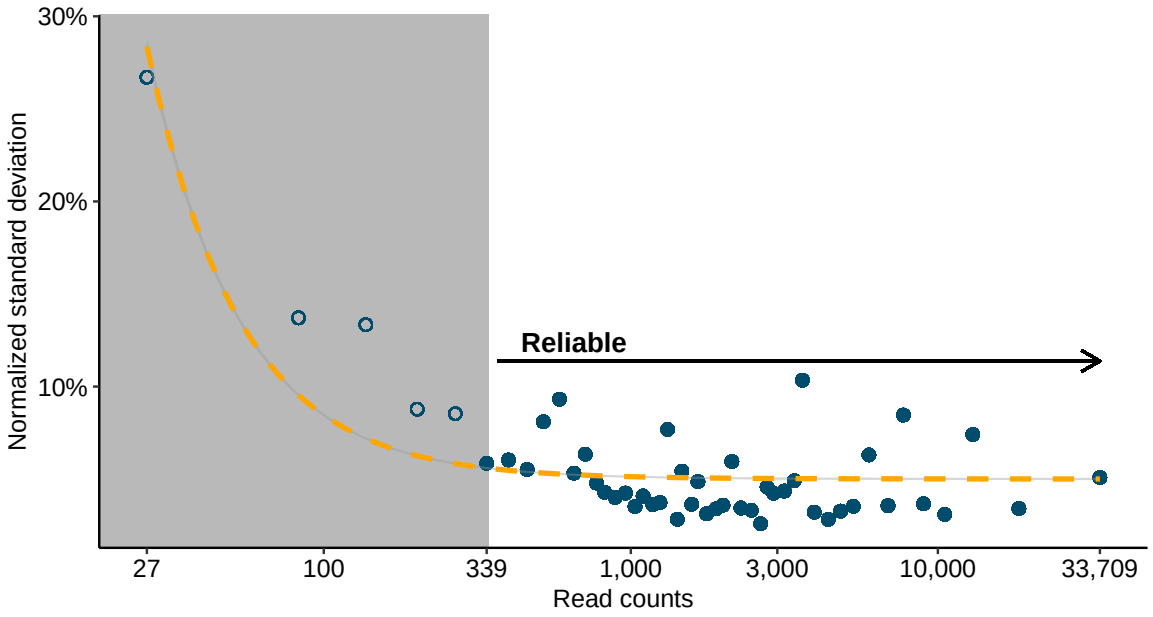

Figure 3.3: **Determining reads per gene thresholds** Biological replicates were used to determine read count thresholds when comparing genetic expression between data sets. Total read counts per gene were calculated by summing reads per gene for each replicate. Genes were binned according to this total 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.

value was used to calculate read count thresholds necessary shown 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 $52\,\mathrm{reads}$ to $573\,\mathrm{reads}$, where samples with greater total reads had greater count thresholds. This criteria filtered approximately $1\,\%$ of total nascent chains calculated per sample.

3.3.2 Translational landscape under heterologous conditions

We used Ribo-seq to survey nascent chain production in GS115 Mut$^+$ cultures collected before methanol induction and 3 and 24 after methanol induction. We

determined how cells differentially express proteins related to cellular processes and signaling, information storage and processing, metabolism, and functions that have yet to be characterized (*Figure. 3.4*). Summed nascent chain production for genes

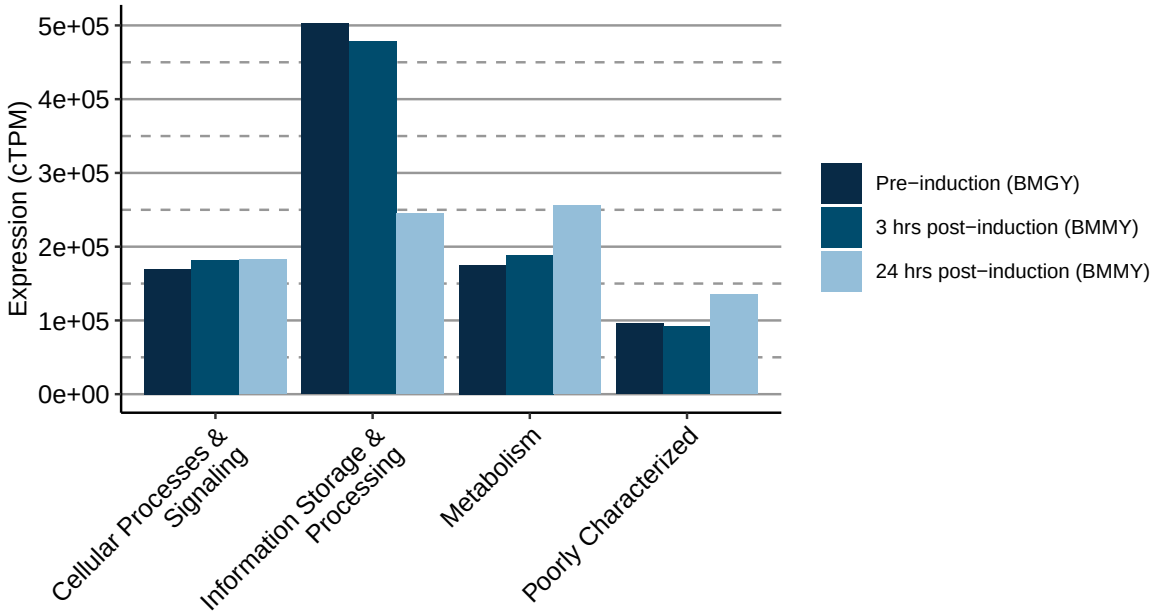

Figure 3.4: **Nascent chains produced under heterologous conditions** Nascent chain production is compared in GS115 Mut^{+} cultured in buffered glycerol media (BMGY), and 3 hours and 24 hours after induction using buffered methanol media (BMMY). The most significant genes differentially expressed after 24 hours are those involved in metabolism and information storage and processing.

involved in cell processes and signaling is relatively conserved over time and is an indication of these functions' vitality. However, some genes involved in UPR like *IRE1* and *HAC1* show 3.2-*fold* and 6.3-*fold* increased expression respectively after 24 hours of methanol induction. As well, the sum of nascent chains produced by genes involved in ERAD show $1.$-*fold* increased expression. From pre-induction to 24 hours after induction, the most significant changes in ontologically categorized nascent chain production occur for those involved in information storage and processing. Differential expression of these genes was predominated by those involved in translation and ribosome biogenesis as their expression decreases by $76\,\%$ (*Figure 3.5*). Decreased expression of ribosomal proteins was accompanied by

increased expression of RNA binding proteins like *LHP1*, where there was nearly a 2-*fold* change. For metabolism, we find that differential expression is increased across all genes but that those involved in the synthesis, transport, and catabolism of secondary metabolites, lipids, and carbohydrates are most affected. While those involved in amino acid transport and biosynthesis are not the most differentially expressed as a whole, we do see differential expression for genes like *GCN4*, 20.6-*fold* increase, and *GLN1*, 7.1-*fold* increase. After 24 hours of induction, the most differentially expressed uncharacterized proteins are those predicted to localize in the peroxisome, where energy production begins for methanol in the methanol utilization (MUT) pathway.

In the MUT pathway, AOX is generated strictly in presence of methanol and absence of glucose. While there are two genes that encode for AOX, the majority of AOX activity in GS115 Mut$^+$ is expressed through *AOX1* as our datasets detect 32-*fold* greater expression from *AOX1* than *AOX2* after 24 hours of growth in methanol based media. AOX is generated in the peroxisome and catalyzes the breakdown of methanol into hydrogen peroxide and formaldehyde. We find 4.2-*fold* increased expression of peroxisomal encoding genes after 24 hour growth in methanol media. In the peroxisome, hydrogen peroxide is degraded into oxygen and water by catalase (*CTA1*) which we see differentially expressed by *130.7-fold*. As hydrogen peroxide causes oxidative stress, we also observed increased expression involved in oxidative stress responses for genes like *YAP1*, 5.3-*fold*, and *GSH2*, 12.6-*fold*. Formaldehyde is assimilated after converting to dihydroxyacetone (DHA) and glyceraldehyde-3-phosphate (GAP) by dihydroxyacetone synthase (*DAS1*, *DAS2*, and possibly *TLK1*). Our datasets show that translation of *DAS2* occurs more extensively than *DAS1* as it produces 1.5-*fold* more nascent chains. Outside of the peroxisome, formaldehyde is dissimilated into formate by formaldehyde

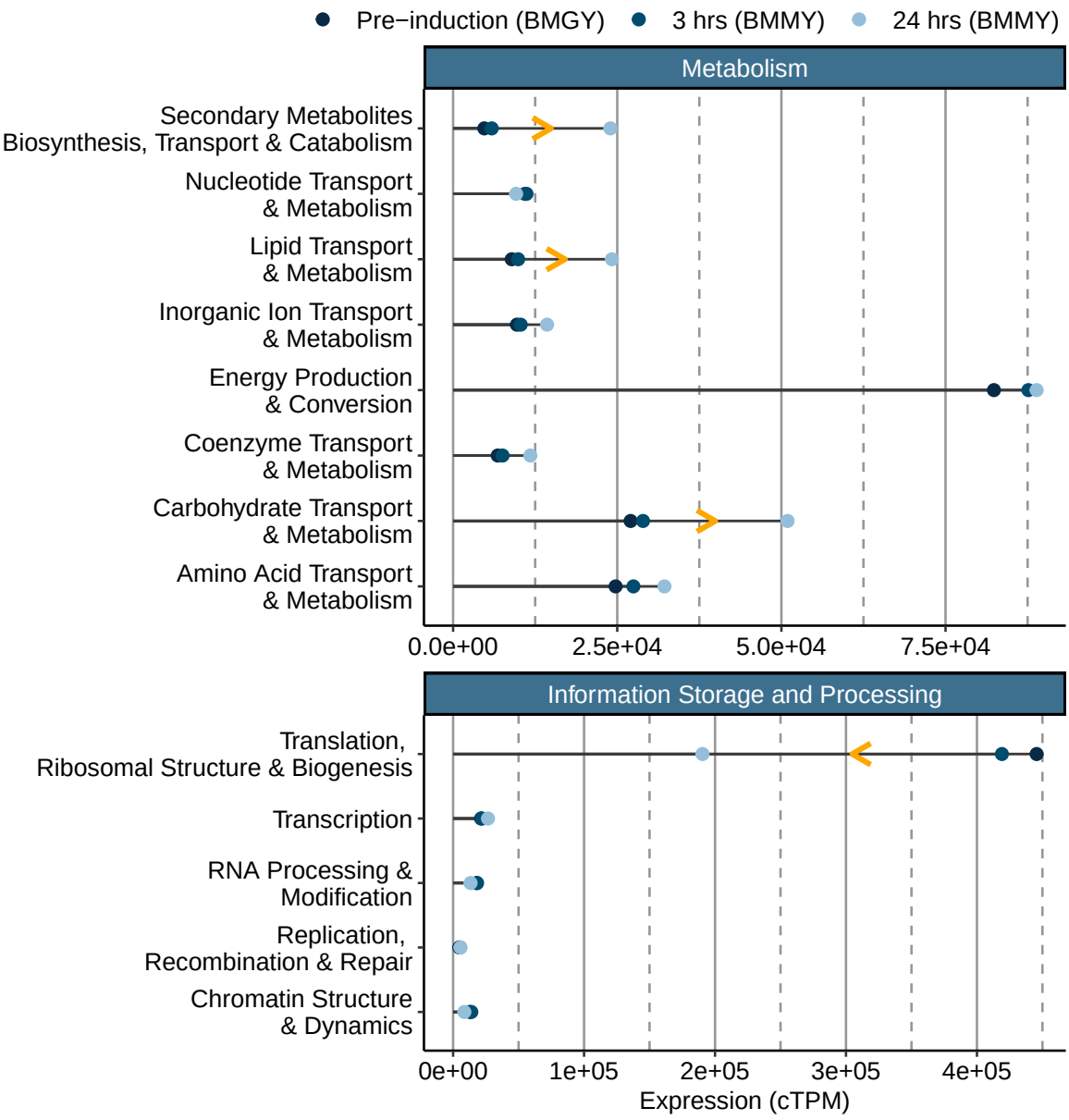

Figure 3.5: **Nascent chains produced under different conditions** Nascent chain production is compared in GS115 Mut$^+$ cultured in buffered glycerol media (BMGY), and 3 hours and 24 hours after induction using buffered methanol media (BMMY). Cells struggle to metabolize methanol after induction and nearly all genes involved in metabolism are positively differentially expressed. As this occurs, expression of genes involved in translation and ribosome biogenesis concomitantly decreased.

dehydrogenase (*FLD*) and carbon dioxide by formate dehydrogenase (*FDH*) for energy production. While we see a 19.3-*fold* increase in summed nascent chain production for all genes involved in the MUT pathway, the greatest increases in

expression are for *FDH*, *AOX1* and *CTA1* in that order.

3.3.3 Heterologous expression and host protein biogenesis demands

We sought to understand how heterologous production affects host protein synthesis by comparing translation 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 (*Figure 3.6*). However, expression diverges significantly over time between the two

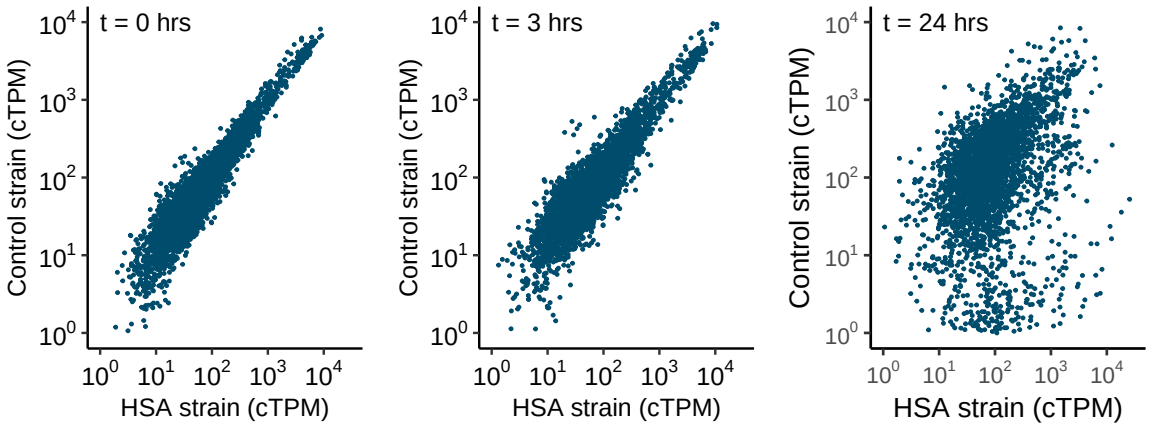

Figure 3.6: **Divergence of translational landscape after heterologous expression** Prior to heterologous induction, nascent proteins produced per gene for GS115 Mut$^+$ and GS115 Mut^S *ALB* are highly correlated and have a Pearson's R of 0.97. After heterologous induction with methanol media, genetic expression diverges between the two strains where they have Pearson's R correlations of 0.94 and 0.32 after 3 and 24 hours.

strains as they show Pearson's correlations of 0.94 and 0.32 after 3 and 24 hours of methanol induction. While both strains showed increased expression of metabolic genes 24 hours after methanol induction, we observe 0.3-*fold* differences across the board. Indeed, those involved in the MUT pathway were both more greatly expressed over time but showed 0.2-*fold* differences in expression between GS115 Mut$^+$ and GS115 Mut^S *ALB*. As methanol utilization produces hydrogen peroxide,

genes associated with oxidative stress like *CTA1* (0.2-*fold*), *GSH2* ($0.$-*fold*), *YAP1* (0.5-*fold*), *GLR1* (0.5-*fold*), and peroxisomal proteins in general (0.3-*fold*) were all less expressed in GS115 Mut^{S} *ALB* than they were in GS115 Mut^{+}. The relative stoichiometry of MUT pathway proteins were not largely changed, as GS115 Mut^{S} *ALB*'s *DAS1* and *DAS2* showed a similar 1.5-*fold* difference in their relative expression ratios as previously observed in GS115 Mut^{+}. Trends of increased expression over time but decreased expression of GS115 Mut^{S} *ALB* compared to GS115 Mut^{+} were also shown for the UPR gene *HAC1* ($0.$-*fold* difference), summed genes involved in ERAD ($0.$-*fold* difference), and the amino acid biosynthesis genes *GCN4* and *GLN1* ($0.$-*fold* differences). However, some genes were opposite this trend. Those involved in ribosome biosynthesis were negatively expressed over time in both strains but showed 1.8-*fold* differences between strains. Grouped by their predicted localizations, extracellular proteins showed 3.5-*fold* differential expression between GS115 Mut^{+} and GS115 Mut^{S} *ALB* and represented the most changed group. A majority of these proteins, $58\,\%$, are involved in cell wall biogenesis or have unknown function but have been previously speculated to be incorporated into the cell wall.[136] Cell wall biogenesis may challenge Sec-translocon availability as we also observe 3.9-*fold* increased expression for genes encoding translocon subunits between strains.

Heterologous proteins traffic through Sec-translocons co-translationally while host cell proteins may do so using co- or post-translational pathways. Each pathway requires distinct translocons, and these translocons are composed of partially overlapping sets of subunits, so we were interested in native proteins that use each. We estimate 56 protein products to enter the ER post-translationally and 931 protein products to enter the ER co-translationally. Before induction, approximately $13\,\%$ of nascent chains produced by each strain were predicted to enter the secretory pathway. Of these proteins, there was a relatively equal amount that

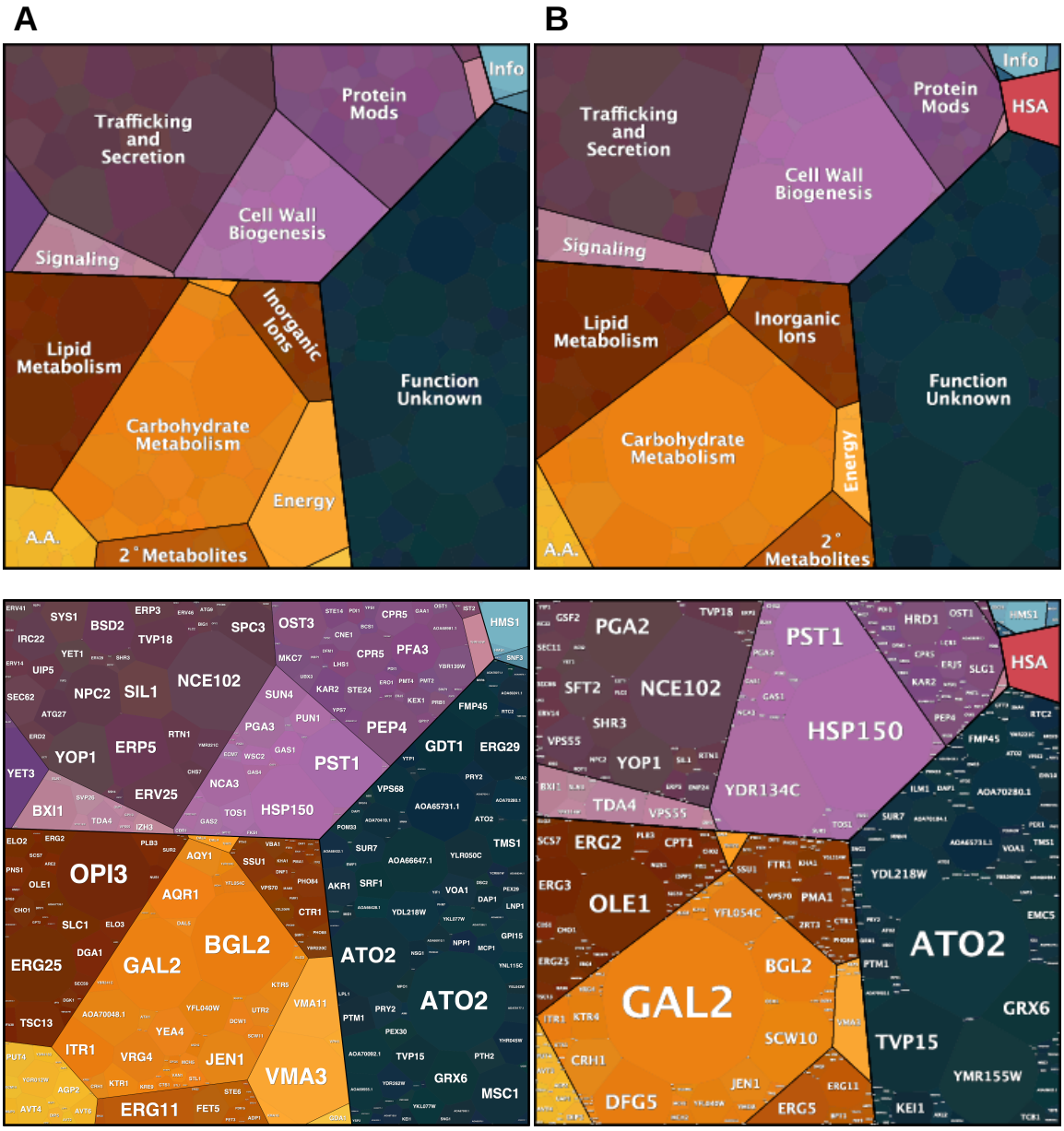

Figure 3.7: **Co-translational flux through the ER in GS115 Mut$^+$ and GS115 Mut^S *ALB* 24 hours after induction** Non-mitochondrial proteins are predicted to enter the secretory pathway co-translationally if they have greater than log_2 membrane enrichment in YPD studies. Gene products are grouped 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. ***a*** Co-translational trafficking in GS115 Mut$^+$. ***b*** Co-translational trafficking in GS115 Mut^S *ALB*.

entered the ER co- and post-translationally. After 24 hours of methanol induction, each strain produced a greater but relatively equivalent amount of the nascent

chains predicted to enter the early secretory pathway, approximately $16\,\%$. However, the ratio of nascent chains entering the ER co- and post-translationally greatly diverges between strains as $32\,\%$ were predicted to enter post-translationally for GS115 Mut^{S} *ALB* while only $5\,\%$ were predicted for GS115 Mut^{+}. Extracellular and membrane proteins represented the majority, $92\,\%$, of the 6.6-*fold* difference in nascent chains predicted to enter the ER post-translationally between strains. While GS115 Mut^{S} *ALB* and GS115 Mut^{+} both show approximately 2-*fold* increased expression of co-translationally translocated proteins after 24 hours, the genes that are differentially expressed between the two strains appear to be much different (*Figure 3.7*). The most differentially expressed of these proteins are *HRD3*, involved in ERAD, and *LAS21*, involved in GPI synthesis, as they show 54-*fold* and 3-*fold* increased expression in GS115 Mut^{S} *ALB* than GS115 Mut^{+}.

We asked which host cell proteins might limit entrance of heterologous proteins into the ER by sequestering the most Sec-translocons at different stages of heterologous expression in GS115 Mut^{S} *ALB*. In generating a "hit list" of host cell proteins, we were interested in those that may limit bioproduction so we excluded ER-resident proteins as their deletion potentially harms folding and secretion of heterologous proteins. The variety and difference in levels of nascent chains entering the secretory pathway after methanol induction are greater for those entering co-translationally than those entering post-translationally (*Table 3.2*). 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. Before methanol induction, the most highly expressed co-translationally translocated proteins are *PST1*, *PMA1* and *BGL2*. *PST1* and *BGL2* are non-essential secreted proteins involved in cell wall maintenance. *PMA1* is a long-lived essential membrane protein involved regulating cytoplasmic pH. After methanol induction, the most highly expressed

Table 3.2: **Host cell proteins that sequester the most Sec-translocons in GS115 Mut^S *ALB***

Gene	Description	Classification	Nascent chains (cTPM)
Pre-induction			
Co-translationally translocated [a]			
PST1	Cell wall protein	Secreted [b]	1,676.0
PMA1	Plasma membrane H+-ATPase	Membrane [c]	1,103.0
BGL2	Endo-beta-1,3-glucanase	Secreted	1,057.0
Post-translationally translocated [d]			
SCV12161.1	Uncharacterized protein	Secreted	12,800.0
XP_002494332.1	Uncharacterized protein	Secreted	12,292.0
SPI1	GPI-anchored cell wall protein	Secreted	11,689.0
24 hour post-induction			
Co-translationally translocated			
GAL2	Low affinity glucose transporter	Membrane	6,304.0
YDR134C	Putative lectin-like protein	Secreted	5,366.0
ADY2	Acetate transporter (isoform 4)	Membrane	4,431.0
Post-translationally translocated			
SPI1	GPI-anchored cell wall protein	Secreted	12,469.0
XP_002494332.1	Uncharacterized protein	Secreted	12,154.0
SCV12161.1	Uncharacterized protein	Secreted	11,447.0

[a] Membrane enriched
[b] Contains signal sequence and one or lesser transmembrane domains (TMD)
[c] Contains signal sequence and more than one TMD or no signal sequence and one TMD
[d] Not membrane enriched, contains signal sequence or C-terminal GPI anchor

co-translationally translocated proteins are *GAL2*, *YDR134C*, and *ADY2*. *GAL2* and *ADY2* are non-essential membrane proteins involved in carbohydrate import and acetate transport respectively. *YDR134C* is a non-essential secreted protein involved in cell wall maintenance that is homologous to *S. cerevisiae*'s paralog of *CCW12*. For nascent chains that enter the ER post-translationally, the most highly expressed proteins are secreted and remain conserved before and after induction and are *SPI1*, *XP_002494332.1*, and *SCV12161.1*. These proteins are relatively small and are likely non-essential cell wall constituents as they are predicted to localize extracellularily.

3.4 Discussion

The yields of heterologous protein products in yeasts often suffer from bottlenecks in biogenesis.[2,142] Current methods for increasing heterologous production are wide ranging and include optimization of growth conditions[143,144] including optimization of methanol concentration for producing different protein products in *K. phaffii*,[11] modification of mRNA structural elements,[145] engineering signal sequences,[146,147] and modification of genes involved in the secretory pathway.[148–150] While these strategies improve secretion, increases in production titers are incremental and optimizations that work well in one condition may not work well in others.[151,152] We hypothesize that cells' biogenetic machinery has co-evolved under the demands of their own proteome. Understanding how a production chassis' uniquely derived protein expression system operates under heterologous conditions may provide unique insight to improve bioproduction.[153,154] In *K. phaffii*, transcriptomic studies have been used to identify gene targets differentially expressed during heterologous production that may be overexpressed to improve bioproduction;[155] 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 for genes involved in the UPR and not to increased heterologous expression.[48] We aimed to use -omics based approaches so that we may identify bottlenecks that may hinder bioproduction in our conditions.[16,93,156] Indeed, researchers have recently calculated host proteome biosynthesis demands in Chinese hamster ovary cells grown under heterologous conditions for the guided depletion of non-vital mRNA to enhance growth rate, improve product quality, and increase protein secretion.[26,108]

Ribo-seq is a high throughput sequencing technique to measure protein synthesis levels by inferring ribosome abundance at each codon in each transcript.[80] As a metric tool, Ribo-seq more closely correlates with standard proteomics than RNA-seq[32] and is much higher throughput than mass spectrometry while maintaining the ability to accurately predict mature protein stoichiometry.[35,116] The utility of Ribo-seq is not without inherent complications. Isolation of ribosome protected mRNA footprints requires rRNA subtraction techniques that are multitudinous in their variety.[36,139,157,158] We present an rRNA subtraction technique that utilizes commercially available depletion agents[159] and probe-directed degradation with DSN.[158] This rRNA subtraction pipeline shows great success compared to yeast studies using other strategies.[37,160] However, the time required for sterile-filtration before flash freezing varied between samples as they were collected at different culture densities and with different amounts of secreted proteins in their media. Ribosome integrity likely also varied as rRNA subtraction was more efficacious for samples with filtration times and culture densities similar to the sample that the depletion probes were designed against.[39] Variable ORF sequencing depth between samples is consequential to non-uniform rRNA subtractions and complicates differential expression analysis.

Common differential expression tools like DESeq2 and edgeR normalize read counts without considering transcript length and assume that most genes are not differentially expressed between biological replicates.[161,162] This is not ideal for complex Ribo-seq studies with few or no replicates that aim to quantify nascent polypeptide chain synthesis and their biogenesis demands. Given that most genes are translated at similar rates,[116] calculating nascent chain synthesis requires transcript length considerations as shorter transcripts may express more nascent chains than longer transcripts given the same time constraints and ribosome availability.

The number of nascent chains dictates levels of sequestration for biogenesis factors used in the secretion pathway like GPI-anchors,[121] receptor-mediated transport proteins,[122,123] and translocatory ratcheting proteins located near Sec-translocons.[114] Transcript length corrections are not trivial, however, as Ribo-seq library preparation collects ribosomes unevenly along transcripts.[79,91] This is ameliorated at the codon level, where Ribo-seq reads are scaled using metagene corrections to computationally distribute them more evenly along transcripts[35,136] to better correlate with protein abundance.[38] We present a novel method for calculating read count thresholds in differential expression analyses that does not require biological replicate for each sample, though at least three biological replicates are needed, and uses metagene-scaled reads. This method utilizes a quantile regression similar to other normalization techniques[163,164] and relies on two assumptions: inter-replicate variability is asymptotic at higher read counts[38] and that inter-replicate variability is linearly related to sequencing depth.[165,166] Intra- and inter-sample comparisons were made after normalizing for transcript length and sequencing depth using a modified form of the transcripts per million (cTPM) metric.[136]

In *K. phaffii*, industrial bioproduction typically relies glycerol-based media for cell growth and methanol-based media for heterologous induction.[47,131,132] We compared host protein synthesis between GS115 Mut^{+} and GS115 Mut^{S} *ALB* under these conditions. While both strains are histidine auxotrophs, GS115 Mut^{S} *ALB* is complemented with *Saccharomyces cerevisiae* derived *HIS4*. Differential expression between strains in this regard, however, is not likely significant as glycerol- and methanol-based medias used in this study are histidine sufficient. 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 AOX generated by *AOX1* and/or *AOX2*. In Mut^{S} strains,

AOX production relies solely on *AOX2* expression as *AOX1* is disrupted. Therefor, growth in methanol media is slower for Mut^S than Mut^+ as AOX is produced solely from *AOX2*, which is expressed to a lesser extent than *AOX1*.[15] Many heterologous proteins are ideally expressed and glycosylated using the *AOX1* promoter.[167] For these proteins, higher production titers are observed in the slow growing Mut^S strain than the fast growing Mut^+ strain.[168] Observing protein synthesis of both strains under these conditions provides insight into possible strategies for strain engineering.

During the first step in the MUT pathway, peroxisomal AOX generates high levels of H_2O_2 during methanol catalysis. We find that methanol metabolism leads leads to increased expression of *YAP1* and *GSH2*, where YAP1p is a required transcription factor for *GSH2* which expresses glutathione in the glutathione redox system.[169] 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.[170,171] While RNA-seq has been used to study oxidative stress responses proceeding methanol metabolism,[172] Ribo-seq is a more sensitive and appropriate tool for quantifying protein levels as oxidative stress increases the frequency of post-transcriptional modifications.[32,173] For instance, RNA-seq finds *DAS1* and *DAS2* equally expressed after methanol induction[168] 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.[173] As well, our analyses are congruent with previously observed reductions in protein synthesis rates consequential to oxidative stress[174] as we find decreased expression of genes encoding ribosome proteins and increased expression of genes

encoding RNA-binding proteins thought to stabilize slowly translating transcripts from degradation.[175] 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. Compared to GS115 Mut$^+$, GS115 Mut^S *ALB* also shows lower overall expression levels for genes involved in the UPR and ERAD. As heterologous production results in greater expression of these genes,[48] and only GS115 Mut^S *ALB* expresses heterologous proteins between the two strains, the potential ramifications that methanol induced oxidative stresses have on bioproduction can therefor be seen as significant. Engineering oxidative stress response pathways in *K. phaffii* is appropriate for increasing methanol induced bioproduction. Indeed, overexpressing stress response proteins has been previously been shown to lower the UPR response while increasing secretion.[154,169]

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,[176] N-terminal signal sequences,[99] and/or glycosylphosphatidylinositol (GPI) anchors.[95] SRP is often pre-recruited to the ribosome nascent chain complex (RNC)[177,178] and thus binds quickly to an emerging hydrophobic targeting sequence.[21] Some proteins do not utilize hydrophobic targeting domains for co-translational translocation and are guided to the ER using mechanisms involved the 3'UTR of their protein encoding transcripts.[179] Some proteins containing N-terminal signal sequences complete translation before they

have time to reach the ER[180] and are instead translocated post-translationally.[181] These proteins typically contain few amino acids.[21] 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.[95] 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[136] 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[105] and that proteins' ER translocation routes are contingent on their sequence features and constituent number of amino acids.[21,136]

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.[182] The outer of layer of the cell wall consists of glycosylated mannoproteins covalently linked to the β 1,3-glucan-chitin network directly or disulfide bound to other cell wall proteins. Cell wall mannoproteins affect stability and resistance to stress.[183–185] As the extracellular and membrane proteins that largely constitute differences between strains are not those shown to be inductively expressed from oxidation,[186] it would appear that reorganization of GS115 Mut^S *ALB*'s cell wall is instead consequent to stresses imposed by heterologous secretion. Therefor, 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[136,187,188] in *K. phaffii*. As such, the signal sequence of *SPI1*'s paralog, *SED1*, has successfully been used to increase secretion of β -glucosidase and endoglucanase-II.[189] Before induction, the most highly expressed co-translationally translocated gene products are from *PST1*, *PMA1*, and *BGL2*. Interestingly, overexpression of *RPP0* has been shown to increase heterologous secretion by mechanisms suspected to decrease *PMA1* expression.[190] After induction, the most highly expressed co-translationally translocated gene products are from *GAL2*, *YDR134C*, and *ADY2*. As galactose is preferentially incorporated into cell wall glucan over glucose,[191] we speculate that *GAL2* is differentially expressed secondary to increased overall expression of cell wall mannoproteins. This is particularly advantageous for strain engineering under the heterologous conditions used for this experiment as methanol media does not contain galactose and *GAL2* disruption should not affect cell growth or viability. For *K. phaffii* cultured in methanol media, depletion of available methanol as a carbon substrate induces *ADY2* and the subsequent uptake of lactic acid for energy production.[192,193] This may indicate that GS115 Mut^S *ALB* may benefit from greater than $0.5\,\%$ methanol concentrations or more frequent than once daily methanol spikes for greater bioproduction. Previously, disruption of the cell wall mannoprotein encoding gene, *CWP2*, increased heterologous secretion coincident to increased expression of genes involved in ribosome biogenesis and decreased expression of cell wall protein encoding genes like *SED1* and *CCW12* (paralogous to *YDR134C* in *S. cerevisiae*).[194] While future studies are required to elucidate the mechanisms, the cell wall mannoprotein encoding genes *SPI1* and *YDR134C* are viable targets for

improving HSA secretion in *K. phaffii* cultured in methanol media.

3.5 Conclusions

Heterologous production is a complex phenomena that requires translational and secretory machinery that is limited in number and processivity. Our analysis in *K. phaffii* reveals invaluable insights into these conditions that are useful for process control and strain engineering. First, Ribo-seq is a powerful tool for surveying host proteome demands and requires semi-specific rRNA reduction strategies inherent to different conditions. Second, heterologous conditions involving methanol induction require tightly regulated levels of substrates and can be further improved. Third, host protein flux through the ER change in response to heterologous protein production. Finally, the variety and levels of host proteins entering the secretory pathway are unique to different stages of heterologous expression.

Conclusion

Therapeutic biologics are enormously useful and the scope of their utility continues to grow. However, biologics are among the most expensive treatment options available today. Expenses associated with producing biologics trickle down to patients in magnitudes that become prohibitive in developed and especially developing societies. As such, a great deal of progress has been made to improve our understanding of biogenesis networks so that production chassis can be more efficiently levied to produce proteins with greater yields. The majority of these efforts have been made in model organisms grown under conditions not utilized for heterologous protein production. This is problematic, as the range of tractable species used as microbial cell factories continues to expand. Organisms' uniquely derived proteomes evolved under the demands of their ecological niche, and strain modifications that improve bioproduction in one organism do not transfer in their efficacy to another. Even within the same organism, strain modification that improve bioproduction in one condition are also not necessarily transferable to others. Thus, technologies and strategies to improve our understanding of host expression systems for rational strain engineering needs to be high throughput and adaptable to lesser understood organisms. The work presented in this dissertation utilizes ribosome profiling (Ribo-seq) as a tool to identify rational targets in *Komagataella phaffii* for increased bioproduction.

In chapter one, we develop a Ribo-seq pipeline for quantifying protein synthesis in non-model organisms that includes library preparation, genome annotation corrections, computation masking, and analysis. While the method for preparing Ribo-seq libraries was certainly not pioneered in this study, several adaptations were made to an existing library preparation protocol from Nicholas Ingolia's research group[37] to include more recent advancements and findings in nucleic acid research. These included modifications to cell lysis, sucrose gradient analysis for optimization of nuclease digestion, ligation of adenylated linkers, reverse transcription, and preliminary read processing. The library preparation strategy yielded libraries with great fidelity and revealed that Illumina's Ribo-Zero Gold kit for rRNA reduction designed for *S. cerevisiae* was sufficient for reductions in *K. phaffii* collected under the conditions used in this chapter. However, it is worth noting that Illumina's Ribo-Zero Gold kit for rRNA reduction is not effective for all ascomycetes. While it was not discussed in this dissertation, we also used the described library preparation strategy in *Yarrowia lipolytica*, *Trichoderma reseii*, and *Ogatae polymorpha*. For these organisms, rRNA contamination was significant (greater than $95\,\%$) and likely a result of lower degrees of homology with *S. cerevisiae* and/or ribosome translocation differences leading to nuclease generated rRNA fragments that escape Illumina's reduction kit. Indeed, this is the most limiting factor in Ribo-seq non-model organisms. We recommend using our library preparation protocol for one sample followed by sequencing with lesser read depth than HiSeq4000 or NextSeq (Illumina sequencing instruments used for this dissertation) using Illumina MiSeq. Sequenced reads may then be used for designing complementary probes for targeted depletion with DSN as discussed in chapter three.

After development of the library preparation protocol, we found many loci where Ribo-seq derived translational start and stop sites disagreed with prior an-

notations. Since the length of a transcript is a critical parameter in interpreting Ribo-Seq, improperly annotated translational start and stop sites lead to inaccurate protein synthesis calculations. Evidence based modeling is ideally used for generating genome annotations as *de novo* gene predictors often misannotate open reading frame boundaries and do not recognize very small genes, especially for organisms like *K. phaffii* with tightly packed genomes. However, previous studies in *K. phaffii* were previously limited to genome and RNA sequencing. We used Ribo-seq (evidence for demarcating open reading frames) and long read RNA sequencing (evidence for demarcating untranslated regions) to improve *K. phaffii* GS115's genome annotation. This new genome annotation led to better protein sequence feature predictions, expanded previous genome annotations by several hundred genes, and have been made publicly available on NCBI for other researchers to use. As such, this strategy would also greatly benefit research for other non-model organisms where genome annotations are curated using lesser sequencing based evidence than *K. phaffii*'s previously was.

Ribo-seq reads are small and are susceptible to mapping to multiple locations in *K. phaffii*'s tightly packed genome. We improved upon previous strategies to computationally mask individuals codons that map to multiple locations in the genome. This reduced biases inherent to canonical methods that either discard or randomly assigned mapped reads. However, as Ribo-seq disproportionately maps reads at the beginning of transcripts, masked codons located at the beginning of a gene will disproportionately affect gene read counts than masked codons at the end of a gene. We developed a metagene correction strategy to computationally distribute reads more evenly along transcripts and ameliorate for codon mask location biases. In doing so, we used metagene correct read counts and the ratio of masked codons per gene to create novel metrics to quantify nascent chain production and ribosome

sequestration. While ribosome distribution biases are widely observed in Ribo-seq experiments, we observe varying degrees of distribution imbalances contingent on the library preparation technique used. Ribo-seq experiments would benefit from a formal comparison of current approaches' effects on ribosome distributions along transcripts. An optimized strategy would ideally capture ribosome distribution imbalances secondary only to ribosome stalling and elongation rate changes. This would greatly improve our understanding of how different conditions affect genes' translational dynamics.

In chapter two, we use the Ribo-seq pipeline from chapter one with subcellular fractionation to quantify demands on co- and post-translational ER translocation pathways. Cells are a crowded environment comprised of thousands of endogenous proteins. Therefor, isolating a heterologous protein is greatly simplified (in both procedural complexity and cost) if it is secreted into the host media. However, entry into the early secretory pathway has been shown to be the rate limiting bottleneck in heterologous secretion. Subcellular fractionation allowed us to capture actively translating ribosomes in the cytosol and on the membrane of the endoplasmic reticulum. We used the log_2 ratio of Ribo-seq reads from membrane and cytosolic samples cultured in YPD to calculate membrane enrichment scores for each gene product; these enrichment scores were used to categorize those co-translationally translocated. As well, genes that were predicted to enter the secretory pathway based on protein sequence features but failed to do so were categorized as post-translationally translocated. Highly expressed proteins that enter the endoplasmic reticulum through these pathways sequester biogenesis machinery limited in number and processivity. Host sequestration complicates heterologous translocation into the endoplasmic reticulum and is therefor a significant contributor to the early secretory bottleneck.

For *K. phaffii* and *S. cerevisiae*, we found that a small number of host genes are responsible for most of the proteins entering the secretory pathway. Of these genes, GPI-anchored protein components of the cell wall represent the greatest number of nascent chains within the secretory pathway. We also show that co-translational translocation pathways in general must accommodate a wider variety of proteins than post-translational pathways but that the number of proteins entering each pathway is approximately equal. In both strains, a protein's propensity to translocate into the endoplasmic reticulum co-translationally was show to be contingent on the number of its constituent amino acids. In contrasting *K. phaffii* and *S. cerevisiae*, we show that orthologs may enter the endoplasmic reticulum through different translocation pathways. As well, the most highly expressed genes whose products enter the endoplasmic reticulum co- and post-translationally was distinct for each organism. Thus, we show that while there are similarities in the expense of biogenetic resources needed to translocate proteins through the endoplasmic reticulum, there are distinct differences in the set of proteins that sequester the most of those resources between organisms. These differences help to explain why host modifications that are effective for increasing secretion in *S. cerevisiae* are not as effective in *K. phaffii*. Future work would benefit by comparing translocation pathways for genes of the same strain grown under different conditions. Elucidating differences in translocation pathways for same genes would greatly increase our understanding of endoplasmic reticulum targeting mechanisms.

In chapter three, we use the Ribo-seq pipeline from chapter one along with *K. phaffii*'s translocation pathway characterizations from chapter two to characterize host proteome demands of cultures grown under heterologous conditions. In *K. phaffii*, industrial bioproduction typically relies on large scale growth in glycerol media before transferring cells to methanol media for heterologous expression

under the control of the *AOX1* promoter. Two commonly used strains for this purpose are GS115 Mut$^+$ and GS115 Mut^S. The Mut$^+$ strain produces AOX using *AOX1* and *AOX2* whereas the Mut^S strain produces AOX using only *AOX2*, the constitutively less active gene. As such, the Mut$^+$ strain produces more AOX per unit time than Mut^S, thus allowing it to grow more quickly secondary to its higher rate of methanol utilization. Despite this, the Mut^S strain produces greater protein yields than the Mut$^+$ strain. We compared the translatomes of GS115 Mut$^+$ and GS115 Mut^S *ALB* before and after methanol induction. In doing so, we further optimized our Ribo-seq pipeline by incorporating more effective rRNA depletion strategies than those commercially available and by developing a technique requiring minimal biological replicates to determine read count thresholds for differential expression analysis. This allowed us to answer two fundamental questions whose answers provide great insight into identifying gene targets to rationally engineer *K. phaffii* for increased bioproduction. Why does the slower growing Mut^S strain produce greater heterologous yields than its faster growing counter strain? How does production and secretion of heterologous proteins affect host protein synthesis and endoplasmic reticulum trafficking?

In answering our first question, we show that the Mut^S strain may express greater levels of heterologous proteins than the Mut$^+$ strain for reasons other than previously hypothesized explanations that the full force of its *AOX1* promoter is put solely towards heterologous expression. As Mut$^+$ strains generate more AOX than Mut^S strains, and AOX catalyzes methanol into formaldehyde and H_2O_2, we found that the Mut$^+$ strain also showed greater signs of oxidative stress than the Mut^S strain. This occurred for nearly every oxidative stress marker we could detect using Ribo-seq including translation initiation discrepancies, slower translation elongation rates, increased expression for genes involved in oxidative stress re-

sponses, the unfolded protein response, and endoplasmic reticulum associated degradation, and significantly decreased expression of genes involves in translation and ribosome biogenesis. While these markers were also observed in Mut^S, the significant difference in their intensity compared to the Mut^+ strain begins to explain differences in their known production titers. From these findings, we recommend overexpressing genes involved in handling oxidative stress (such as those involved in the glutathione redox system) for heterologously producing *K. phaffii* cultured in methanol conditions. As well, our findings also indicated that *K. phaffii* Mut^S strains would benefit from greater methanol concentrations in their growth media than Invitrogen recommends in their catalog as they also differentially expressed genes indicative of carbon substrate deficits after 24 hours of growth in methanol media. In answering our second question, we find protein components of the cell wall to represent the greatest number of nascent chains entering the endoplasmic reticulum for both strains. However, flux through the endoplasmic reticulum was greater for GS115 Mut^S Albumin than GS115 Mut^+. Additionally, highly expressed cell-wall components entering the endoplasmic reticulum before induction were distinct from the most highly expressed cell-wall components entering the endoplasmic reticulum after induction. Our analysis indicated genes for rational strain engineering that would have not been predicted using other methods.

The methodologies described in this dissertation offer an optimized and logical solution to identify targets for rational strain engineering. Immediate endeavors involve application of insights described herein to increase heterologous protein production in *K. phaffii*. Future endeavors beyond that lie in using this dissertation as a manual to better understand and improve bioproduction in even lesser understood organisms.

References

1. Sanchez-Garcia, L. *et al.* Recombinant pharmaceuticals from microbial cells: A 2015 update. *Microb. Cell Fact.* **15**, 33 (2016).

2. Wang, G., Huang, M. & Nielsen, J. Exploring the potential of saccharomyces cerevisiae for biopharmaceutical protein production. *Curr. Opin. Biotechnol.* **48**, 77–84 (2017).

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

4. Kim, H., Yoo, S. J. & Kang, H. A. Yeast synthetic biology for the production of recombinant therapeutic proteins. *FEMS Yeast Res.* **15**, 1–16 (2015).

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 (2017).

6. Lopes, H. & Rocha, I. Genome-scale modeling of yeast: Chronology, applications and critical perspectives. *FEMS Yeast Res.* **17**, (2017).

7. Cai, P., Gao, J. & Zhou, Y. CRISPR-mediated genome editing in non-conventional yeasts for biotechnological applications. *Microb. Cell Fact.* **18**, 63 (2019).

8. Yamada, Y., Matsuda, M., Maeda, K. & Mikata, K. The phylogenetic relationships of methanol-assimilating yeasts based on the partial sequences of 18S and 26S ribosomal RNAs: The proposal of komagataella gen. Nov. (saccharomycetaceae). *Biosci. Biotechnol. Biochem.* **59**, 439–444 (1995).

9. Kurtzman, C. P. Description of komagataella phaffii sp. Nov. And the transfer of pichia pseudopastoris to the methylotrophic yeast genus komagataella. *Int. J. Syst. Evol. Microbiol.* **55**, 973–976 (2005).

10. Kurtzman, C. P. Biotechnological strains of komagataella (pichia) pastoris are komagataella phaffii as determined from multigene sequence analysis. *J. Ind. Microbiol. Biotechnol.* **36**, 1435–1438 (2009).

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. 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).

13. *Pichia expression kit*. (Life Technologies, 2014).

14. Zahrl, R. J., Peña, D. A., Mattanovich, D. & Gasser, B. Systems biotechnology for protein production in pichia pastoris. *FEMS Yeast Res.* **17**, (2017).

15. Fischer, J. E. & Glieder, A. Current advances in engineering tools for pichia pastoris. *Curr. Opin. Biotechnol.* **59**, 175–181 (2019).

16. 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).

17. Jiang, H. *et al.* Challenging the workhorse: Comparative analysis of eukaryotic micro-organisms for expressing monoclonal antibodies. *Biotechnol. Bioeng.* **116**, 1449–1462 (2019).

18. Crowell, L. E. *et al.* On-demand manufacturing of clinical-quality biopharmaceuticals. *Nat. Biotechnol.* (2018).

19. Love, K. R. *et al.* Systematic Single-Cell analysis of pichia pastoris reveals secretory capacity limits productivity. *PLoS One* **7**, e37915 (2012).

20. 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).

21. Justin W. Chartron, K. C. L. H. &. J. F. Cotranslational signal independent SRP preloading during membrane targeting. *Nature* **536**, 224–228 (2016).

22. 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).

23. Jan, C. H., Williams, C. C. & Weissman, J. S. Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science* **346**, 1257521 (2014).

24. Hessa, T. *et al.* Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature* **433**, 377–381 (2005).

25. Wu, X., Cabanos, C. & Rapoport, T. A. Structure of the post-translational protein translocation machinery of the ER membrane. *Nature* **566**, 136–139 (2019).

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

27. Mori, A. *et al.* Signal peptide optimization tool for the secretion of recombinant protein from saccharomyces cerevisiae. *J. Biosci. Bioeng.* **120**, 518–525 (2015).

28. Sumi, A. *et al.* Purification of recombinant human serum albumin efficient purification using STREAMLINE. *Bioseparation* **8**, 195–200 (1999).

29. Potgieter, T. I. *et al.* Production of monoclonal antibodies by glycoengineered pichia pastoris. *J. Biotechnol.* **139**, 318–325 (2009).

30. Fernandes, L. D., Moura, A. P. S. de & Ciandrini, L. Gene length as a regulator for ribosome recruitment and protein synthesis: Theoretical insights. *Sci. Rep.* **7**, 17409 (2017).

31. Englander, S. W. & Mayne, L. The nature of protein folding pathways. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 15873–15880 (2014).

32. 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).

33. Lareau, L. F., Hite, D. H., Hogan, G. J. & Brown, P. O. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife* **3**, e01257 (2014).

34. Tzani, I. *et al.* Understanding biopharmaceutical production at single nucleotide resolution using ribosome footprint profiling. *Curr. Opin. Biotechnol.* **53**, 182–190 (2018).

35. Taggart, J. C. & Li, G.-W. Production of Protein-Complex components is stoichiometric and lacks general feedback regulation in eukaryotes | elsevier enhanced reader.

36. 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).

37. McGlincy, N. J. & Ingolia, N. T. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* (2017).

38. 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).

39. Gerashchenko, M. V. & Gladyshev, V. N. Ribonuclease selection for ribosome profiling. *Nucleic Acids Res.* **45**, e6 (2017).

40. Kraus, A. J., Brink, B. G. & Siegel, T. N. Efficient and specific oligo-based depletion of rRNA. *Sci. Rep.* **9**, 12281 (2019).

41. Petrova, O. E., Garcia-Alcalde, F., Zampaloni, C. & Sauer, K. Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. *Sci. Rep.* **7**, 41114 (2017).

42. Thompson, M. K., Kiourlappou, M. & Davis, I. Ribo-Pop: Simple, cost-effective, and widely applicable ribosomal RNA depletion. *RNA* **26**, 1731–1742 (2020).

43. Halpin, J. C., Jangi, R. & Street, T. O. Multimapping confounds ribosome profiling analysis: A case-study of the Hsp90 molecular chaperone. *Proteins* **88**, 57–68 (2020).

44. Brar, G. A. & Weissman, J. S. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat. Rev. Mol. Cell Biol.* **16**, 651–664 (2015).

45. Mohammad, F., Green, R. & Buskirk, A. R. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *Elife* **8**, (2019).

46. Love, K. R. *et al.* Comparative genomics and transcriptomics of pichia pastoris. *BMC Genomics* **17**, 550 (2016).

47. 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).

48. Gasser, B. *et al.* Monitoring of transcriptional regulation in pichia pastoris under protein production conditions. *BMC Genomics* **8**, 179 (2007).

49. Lee, S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E2424–32 (2012).

50. Hsu, P. Y. *et al.* Super-resolution ribosome profiling reveals unannotated translation events in arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E7126–E7135 (2016).

51. Ndah, E. *et al.* REPARATION: Ribosome profiling assisted (re-)annotation of bacterial genomes. *Nucleic Acids Res.* **45**, e168 (2017).

52. Laver, T. *et al.* Assessing the performance of the oxford nanopore technologies MinION. *Biomol Detect Quantif* **3**, 1–8 (2015).

53. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The oxford nanopore MinION: Delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17**, 239 (2016).

54. Lu, H., Giordano, F. & Ning, Z. Oxford nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* **14**, 265–279 (2016).

55. Jain, M. *et al.* Improved data analysis for the MinION nanopore sequencer. *Nat. Methods* **12**, 351–356 (2015).

56. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

57. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

58. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

59. Haas, B. J. *et al.* Improving the arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).

60. Haas, B. & Papanicolaou, A. TransDecoder (find coding regions within transcripts). Github. 3.1. *https://github.com › blob › master › index.asciidochttps://github.com › blob › master › index.asciidoc* (2016).

61. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).

62. Testa, A. C., Hane, J. K., Ellwood, S. R. & Oliver, R. P. CodingQuarry: Highly accurate hidden markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics* **16**, 170 (2015).

63. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).

64. Palmer, J. & Stajich, J. Nextgenusfs/funannotate: Funannotate v1.5.3. (2019).

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

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

67. Li, H. *et al.* The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

68. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).

69. Popa, A. *et al.* RiboProfiling: A bioconductor package for standard ribo-seq pipeline processing. *F1000Res.* **5**, 1309 (2016).

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

71. Saccharomyces genome database | SGD.

72. Protein BLAST: Search protein databases using a protein query.

73. 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).

74. Liebermeister, W. *et al.* Visual account of protein investment in cellular functions. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8488–8493 (2014).

75. Almagro Armenteros, J. J., Sonderby, C. K., Sonderby, S. K., Nielsen, H. & Winther, O. DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics* **33**, 3387–3395 (2017).

76. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).

77. Tsirigos, K. D., Peters, C., Shu, N., Kall, L. & Elofsson, A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.* **43**, W401–7 (2015).

78. Pierleoni, A., Martelli, P. L. & Casadio, R. PredGPI: A GPI-anchor predictor. *BMC Bioinformatics* **9**, 392 (2008).

79. 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).

80. Ingolia, N. T. Ribosome footprint profiling of translation throughout the genome. *Cell* **165**, 22–33 (2016).

81. Valli, M. *et al.* Curation of the genome annotation of pichia pastoris (komagataella phaffii) CBS7435 from gene level to protein function. *FEMS Yeast Res.* **16**, (2016).

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

83. Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683 (2013).

84. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).

85. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

86. Baudin-Baillieu, A. *et al.* Genome-wide translational changes induced by the prion [PSI+]. *Cell Rep.* **8**, 439–448 (2014).

87. Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835–840 (2010).

88. Xiao, Z., Zou, Q., Liu, Y. & Yang, X. Genome-wide assessment of differential translations with ribosome profiling data. *Nat. Commun.* **7**, 11194 (2016).

89. Gardin, J. *et al.* Measurement of average decoding rates of the 61 sense codons in vivo. *Elife* **3**, (2014).

90. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

91. Ingolia, N. T. Ribosome profiling: New views of translation, from single codons to genome scale. *Nat. Rev. Genet.* **15**, 205–213 (2014).

92. Tuller, T. *et al.* An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354 (2010).

93. 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).

94. Deshaies, R. J., Koch, B. D., Werner-Washburne, M., Craig, E. A. & Schekman, R. A subfamily of stress proteins facilitates translocation of secretory and mitochondrial precursor polypeptides. *Nature* **332**, 800–805 (1988).

95. Ast, T., Cohen, G. & Schuldiner, M. A network of cytosolic factors targets SRP-Independent proteins to the endoplasmic reticulum. *Cell* **152**, 1134–1145 (2013).

96. Aviram, N. & Schuldiner, M. Targeting and translocation of proteins to the endoplasmic reticulum at a glance. *J. Cell Sci.* **130**, 4079–4085 (2017).

97. Keenan, R. J., Freymann, D. M., Stroud, R. M. & Walter, P. The signal recognition particle. *Annu. Rev. Biochem.* **70**, 755–775 (2001).

98. Costa, E. A., Subramanian, K., Nunnari, J. & Weissman, J. S. Defining the physiological role of SRP in protein-targeting efficiency and specificity. *Science* **359**, 689–692 (2018).

99. 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).

100. Shao, S. & Hegde, R. S. Membrane protein insertion at the endoplasmic reticulum. *Annu. Rev. Cell Dev. Biol.* **27**, 25–56 (2011).

101. Metzl-Raz, E. *et al.* Principles of cellular resource allocation revealed by condition-dependent proteome profiling. *Elife* **6**, e28034 (2017).

102. Klepsch, M. M., Persson, J. O. & Gier, J.-W. L. de. Consequences of the overexpression of a eukaryotic membrane protein, the human KDEL receptor, in escherichia coli. *J. Mol. Biol.* **407**, 532–542 (2011).

103. Farkas, Z. *et al.* Hsp70-associated chaperones have a critical role in buffering protein production costs. *Elife* **7**, e29845 (2018).

104. Yang, L., Yurkovich, J. T., King, Z. A. & Palsson, B. O. Modeling the multi-scale mechanisms of macromolecular resource allocation. *Curr. Opin. Microbiol.* **45**, 8–15 (2018).

105. 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).

106. Feizi, A., Osterlund, T., Petranovic, D., Bordel, S. & Nielsen, J. Genome-scale modeling of the protein secretory machinery in yeast. *PLoS One* **8**, e63284 (2013).

107. Gutierrez, J. M. *et al.* Genome-scale reconstructions of the mammalian secretory pathway predict metabolic costs and limitations of protein secretion. *Nat. Commun.* **11**, 1–10 (2020).

108. Kol, S. *et al.* Multiplex secretome engineering enhances recombinant protein production and purity. *Nat. Commun.* **11**, 1908 (2020).

109. Gerashchenko, M. V. & Gladyshev, V. N. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res.* **42**, e134 (2014).

110. Engel, S. R. *et al.* The reference genome sequence of saccharomyces cerevisiae: Then and now. *G3* **4**, 389–398 (2014).

111. Scannell, D. R., Butler, G. & Wolfe, K. H. Yeast genome evolution–the origin of the species. *Yeast* **24**, 929–942 (2007).

112. Ahn, J. *et al.* Translation elongation factor 1-alpha gene from pichia pastoris: Molecular cloning, sequence, and use of its promoter. *Appl. Microbiol. Biotechnol.* **74**, 601–608 (2007).

113. Balchin, D., Hayer-Hartl, M. & Ulrich Hartl, F. In vivo aspects of protein folding and quality control. *Science* **353**, (2016).

114. Matlack, K. E., Misselwitz, B., Plath, K. & Rapoport, T. A. BiP acts as a molecular ratchet during posttranslational transport of prepro-alpha factor across the ER membrane. *Cell* **97**, 553–564 (1999).

115. Brodsky, J. L., Goeckeler, J. & Schekman, R. BiP and Sec63p are required for both co- and posttranslational protein translocation into the yeast endoplasmic reticulum. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 9643–9646 (1995).

116. 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).

117. Duttler, S., Pechmann, S. & Frydman, J. Principles of cotranslational ubiquitination and quality control at the ribosome. *Mol. Cell* **50**, 379–393 (2013).

118. Itskanov, S. & Park, E. Structure of the posttranslational sec protein-translocation channel complex from yeast. *Science* **363**, 84–87 (2019).

119. Alamo, M. del *et al.* Defining the specificity of cotranslationally acting chaperones by systematic analysis of mRNAs associated with ribosome-nascent chain complexes. *PLoS Biol.* **9**, e1001100 (2011).

120. Diehn, M., Eisen, M. B., Botstein, D. & Brown, P. O. Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. *Nat. Genet.* **25**, 58–62 (2000).

121. Mayor, S. & Riezman, H. Sorting GPI-anchored proteins. *Nat. Rev. Mol. Cell Biol.* **5**, 110–120 (2004).

122. Semenza, J. C., Hardwick, K. G., Dean, N. & Pelham, H. R. ERD2, a yeast gene required for the receptor-mediated retrieval of luminal ER proteins from the secretory pathway. *Cell* **61**, 1349–1357 (1990).

123. Geva, Y. & Schuldiner, M. The back and forth of cargo exit from the endoplasmic reticulum. *Curr. Biol.* **24**, R130–6 (2014).

124. Shen, X.-X. *et al.* Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* **175**, 1533–1545.e20 (2018).

125. Freskgard, P.-O. & Urich, E. Antibody therapies in CNS diseases. *Neuropharmacology* **120**, 38–55 (2017).

126. Yu, H. *et al.* Market share and costs of biologic therapies for inflammatory bowel disease in the USA. *Aliment. Pharmacol. Ther.* **47**, 364–370 (2018).

127. Koren, M. J. *et al.* Long-term Low-Density lipoprotein Cholesterol-Lowering efficacy, persistence, and safety of evolocumab in treatment of hypercholesterolemia: Results up to 4 years from the Open-Label OSLER-1 extension study. *JAMA Cardiol* **2**, 598–607 (2017).

128. Xu, L. *et al.* Trispecific broadly neutralizing HIV antibodies mediate potent SHIV protection in macaques. *Science* **358**, 85–90 (2017).

129. Lobs, 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).

130. Tschopp, J. F., Brust, P. F., Cregg, J. M., Stillman, C. A. & Gingeras, T. R. Expression of the lacZ gene from two methanol-regulated promoters in pichia pastoris. *Nucleic Acids Res.* **15**, 3859–3876 (1987).

131. Cereghino, J. L. & Cregg, J. M. Heterologous protein expression in the methylotrophic yeast pichia pastoris. *FEMS Microbiol. Rev.* **24**, 45–66 (2000).

132. 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).

133. Structural Genomics Consortium *et al.* Protein production and purification. *Nat. Methods* **5**, 135–146 (2008).

134. Owczarek, B., Gerszberg, A. & Hnatuszko-Konka, K. A brief reminder of systems of production and Chromatography-Based recovery of recombinant protein biopharmaceuticals. *Biomed Res. Int.* **2019**, 4216060 (2019).

135. Akopian, D., Shen, K., Zhang, X. & Shan, S.-O. Signal recognition particle: An essential protein-targeting machine. *Annu. Rev. Biochem.* **82**, 693–721 (2013).

136. 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).

137. Travers, K. J. *et al.* Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. *Cell* **101**, 249–258 (2000).

138. 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).

139. 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).

140. Cutadapt — cutadapt 2.3 documentation.

141. Samtools.

142. Mattanovich, D., Gasser, B., Hohenblum, H. & Sauer, M. Stress in recombinant protein producing yeasts. *J. Biotechnol.* **113**, 121–135 (2004).

143. Salari, R. & Salari, R. Investigation of the best saccharomyces cerevisiae growth condition. *Electron Physician* **9**, 3592–3597 (2017).

144. 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).

145. Gustafsson, C., Govindarajan, S. & Minshull, J. Codon bias and heterologous protein expression. *Trends Biotechnol.* **22**, 346–353 (2004).

146. Mori, A. *et al.* Signal peptide optimization tool for the secretion of recombinant protein from saccharomyces cerevisiae. *J. Biosci. Bioeng.* **120**, 518–525 (2015).

147. 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).

148. 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).

149. 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).

150. Ruijter, J. C. de, Koskela, E. V. & Frey, A. D. Enhancing antibody folding and secretion by tailoring the saccharomyces cerevisiae endoplasmic reticulum. *Microb. Cell Fact.* **15**, 87 (2016).

151. 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).

152. 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).

153. Kroukamp, H. *et al.* Strain breeding enhanced heterologous cellobiohydrolase secretion by saccharomyces cerevisiae in a protein specific manner. *Biotechnol. J.* **12**, (2017).

154. 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).

155. 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).

156. Vogl, T. & Glieder, A. Regulation of pichia pastoris promoters and its consequences for protein production. *N. Biotechnol.* **30**, 385–404 (2013).

157. Faridani, O. R. *et al.* Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.* **34**, 1264–1266 (2016).

158. 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).

159. 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).

160. 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).

161. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

162. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

163. Hansen, K. D., Irizarry, R. A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).

164. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

165. Evans, C., Hardin, J. & Stoebel, D. M. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* **19**, 776–792 (2018).

166. Zhang, M. J., Ntranos, V. & Tse, D. Determining sequencing depth in a single-cell RNA-seq experiment. *Nat. Commun.* **11**, 774 (2020).

167. 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).

168. Krainer, F. W. *et al.* Recombinant protein expression in pichia pastoris strains with an engineered methanol utilization pathway. *Microb. Cell Fact.* **11**, 22 (2012).

169. 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).

170. Harding, H. P. *et al.* An integrated stress response regulates amino acid metabolism and resistance to oxidative stress. *Mol. Cell* **11**, 619–633 (2003).

171. 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).

172. 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).

173. 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).

174. Shenton, D. *et al.* Global translational responses to oxidative stress impact upon multiple levels of protein synthesis. *J. Biol. Chem.* **281**, 29011–29021 (2006).

175. Vogel, C., Silva, G. M. & Marcotte, E. M. Protein expression regulation under oxidative stress. *Mol. Cell. Proteomics* **10**, M111.009217 (2011).

176. 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).

177. 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).

178. Berkovits, B. D. & Mayr, C. Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* **522**, 363–367 (2015).

179. Loya, A. *et al.* The 3'-UTR mediates the cellular localization of an mRNA encoding a short plasma membrane protein. *RNA* **14**, 1352–1365 (2008).

180. Rapoport, T. A. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature* **450**, 663–669 (2007).

181. Shao, S. & Hegde, R. S. A calmodulin-dependent translocation pathway for small secretory proteins. *Cell* **147**, 1576–1588 (2011).

182. Klis, F. M., Mol, P., Hellingwerf, K. & Brul, S. Dynamics of cell wall structure in saccharomyces cerevisiae. *FEMS Microbiol. Rev.* **26**, 239–256 (2002).

183. 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).

184. 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).

185. Vaart, J. M. van der, 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).

186. 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).

187. 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).

188. 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).

189. Inokuma, K. *et al.* Enhanced cell-surface display and secretory production of cellulolytic enzymes with saccharomyces cerevisiae Sed1 signal peptide. *Biotechnol. Bioeng.* **113**, 2358–2366 (2016).

190. Wentz, A. E. & Shusta, E. V. Enhanced secretion of heterologous proteins from yeast by overexpression of ribosomal subunit RPP0. *Biotechnol. Prog.* **24**, 748–756 (2008).

191. 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).

192. Yamada, R., Ogura, K., Kimoto, Y. & Ogino, H. Toward the construction of a technology platform for chemicals production from methanol: D-lactic acid production from methanol by an engineered yeast pichia pastoris. *World J. Microbiol. Biotechnol.* **35**, 37 (2019).

193. Pacheco, A. *et al.* Lactic acid production in saccharomyces cerevisiae is modulated by expression of the monocarboxylate transporters Jen1 and Ady2. *FEMS Yeast Res.* **12**, 375–381 (2012).

194. 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).

Appendix

Bash processing

```
# Downloading SRA files
fastq-dump SRA_code -O output_directory --gzip

# Demultiplexing and read trimming with CutAdapt
## Adapters
810=NNNNNATCGTAGATCGGAAGAGCAC
811=NNNNNAGCTAAGATCGGAAGAGCAC
812=NNNNNCGTAAAGATCGGAAGAGCAC
813=NNNNNCTAGAAGATCGGAAGAGCAC
814=NNNNNGATCAAGATCGGAAGAGCAC
815=NNNNNGCATAAGATCGGAAGAGCAC
816=NNNNNTAGACAGATCGGAAGAGCAC
817=NNNNNTCTAGAGATCGGAAGAGCAC
universal=CTGTAGGCACCATCAAT

cutadapt --cut 2
  -q 10
  -a sample=$adapter
  --minimum-length 6
  --maximum-length 50
  -o output-{name}.fastq.gz
  input.fastq.gz

# RNA depletion with Hisat2
hisat2 -x rRNA_genomic.fna -U input.fastq.gz -p 8
  --max-intronlen 1000
  --un-gz depleted-output.fastq.gz
```

```
  -S rRNA.fastq.gz

# Creating sequence alignment maps with Hisat2
hisat2 -x genomic-annotation.fna
  -U depleted-output.fastq.gz -p 8 --max-intronlen 1000
  -S depleted-output.sam

# Index and sort sequence alignment maps with SAMtools
samtools view -bS -q 60 depleted-output.sam -@ 8
  > depleted-output.bam

samtools sort -@ 8 depleted-output.bam
  -o depleted-output.sort.bam

# Predicting signal peptide presence using SignalP 5.0
signalp -fasta protein_fasta.fna
  -prefix protein_fasta.signalP

# Initial processing of DeepLoc output
| grep "#" -v | grep "Predicted" -v | grep "Localization"
  -v | grep "Likelihood" -v | grep "Hierarchical" -v |
  grep "Type" -v >

awk 'BEGIN{OFS="\t"; RS=""; FS="\n"}{print $1,$2}'
```

R processing

Ribo-Seq functions

```
naTozeroRle <- function(rleObject) {
  ixNA <- which(is.na(S4Vectors::runValue(rleObject)))

  if(length(ixNA) > 0) {
    S4Vectors::runValue(rleObject)[ixNA] <- 0
    S4Vectors::runLength(rleObject)[ixNA] <- 0
  }
```

```
  return(rleObject)
}

applyShiftFeature <-
  function(transcGRangesList, shiftValue) {
    # Check parameter validity
    # If missing shift value or if shift value does not
    # inherit from numeric class
    if(missing(shiftValue) ||
       !inherits(shiftValue, "numeric")) {
      shiftValue <- 0
      warning("Incorrect shiftValue parameter! No shift is
               performed!\n")
    }

    # Check transcGRangesList if of class GRangesList
    if(!is(transcGRangesList, "GRangesList")) {
      stop(
        paste("transcGRangesList parameter is of class ",
              class(transcGRangesList),
              "instead of GRangesList!/n",
              sep="")
      )
    }

    # Determine transcript width from transcGRangesList
    transcWidth <-
      GenomicFeatures::transcriptWidths(start(
        transcGRangesList),
        end(transcGRangesList))

    # Takes absolute value of shift value
    absShiftVal <- abs(shiftValue)

    # if width of transcript is smaller than the absolute
    # shiftValue eliminate the transcript
    ixSmallTransc <- which(transcWidth <= absShiftVal)
    if(length(ixSmallTransc) > 0) {
      transcGRangeList <- transGRangesList[-ixSmallTransc]
      transWidth <-
        GenomicFeatures::transcriptWidths(
          start(transcGRangesList),
```

```
        end(transcGRangesList))
}

# If the shiftValue is positive, the start of the
#transcript is shifted
if(shiftValue > 0) {
  usefulRangeOnTransc <- cbind(
    startT = rep(absShiftVal + 1,
                 length(transcGRangesList)),
    endT = transcWidth)
}
# Else it is the end of the transcript that we shift
else {
  usefulRangeOnTransc <- cbind(
    startT=1,
    endT=transcWidth - absShiftVal)
}

# Make list of useful ranges using usefulRangeOnTrans
# (shifted values for transcript) across length of
# transcGRangesList
listeUsefulRanges <-
  lapply(seq_len(length(
    transcGRangesList)),
    function(ixTransc){
      usefulRangeOnTransc[ixTransc,
                          1]:usefulRangeOnTransc[
                            ixTransc, 2]
    }
  )

# For the remaining positions in the transcript, make
# 1bp bins of the genomic positions
# TranscriptLocs2refLocs function for converting
# transcript-based locations in to reference- based
# locations.
shiftedTransc <-
  GenomicFeatures::transcriptLocs2refLocs(
    listeUsefulRanges,
    start(transcGRangesList),
    end(transcGRangesList),
    as.character(
```

```
          S4Vectors::runValue(strand(transcGRangesList))),
        decreasing.rank.on.minus.strand=TRUE)

    # Give names for shiftedTransc the names from original
    # transcGRangesList
    names(shiftedTransc) <- names(transcGRangesList)

    return(shiftedTransc)
  }

# countShiftReads
countReads <- function(
  exonGRanges,
  cdsPosTransc,
  alnGRanges,
  originalAln,
  shiftValue,
  motifSize)  {

  if(missing(cdsPosTransc)) {
    stop("Missing cdsPosTransc parameter!\n")
  }

  if(length(exonGRanges) != length(cdsPosTransc)) {
    stop(
      "Different lengths for exonGRanges and
          cdsPosTransc parameters!\n"
    )
  }

  myCondNA <-
    which(is.na(unlist(cdsPosTransc)) |
            is.null(unlist(cdsPosTransc)))
  if(length(myCondNA) > 0) {
    stop("Non-null, non-NA values for the
             cdsPosTransc parameter!\n")
  }

  if(missing(shiftValue) ||
     !inherits(shiftValue, "numeric")) {
    shiftValue <- 0
    warning("Incorrect shiftValue parameter!
```

```
                No shift is performed!\n")
}

if(!is(exonGRanges, "GRangesList")) {
  stop(paste(
    "exonGRanges parameter is of class ",
    class(exonGRanges),
    " instead of GRangesList!\n", sep = ""))
}

if(!is(alnGRanges, "GRanges")) {
  stop(paste(
    "alnGRanges parameter is of class ",
    class(alnGRanges),
    " instead of GRanges!\n", sep = ""))
}

if (missing(motifSize) ||
    !is(motifSize, "numeric") ||
    motifSize %% 1 != 0 ||
    motifSize <= 0 ||
    !(motifSize %in% c(3, 6, 9))) {
  warning("Param motifSize should be an integer!
              Accepted values 3, 6 or 9.
              Default value is 3.\n")
  motifSize <- 3
}

exonGRangesRestrict <- exonGRanges[names(cdsPosTransc)]
if(length(exonGRangesRestrict) <= 5) {
  stop(
    "Less than 5 common transcripts btw exonGRanges
        and cdsPosTransc!\n")
}
else {
  if (length(exonGRangesRestrict) <= 10) {
    warning("Less than 10 common transcripts between
                  exonGRanges and cdsPosTransc!\n")
  }
}

transcWidth <-
```

```
  GenomicFeatures::transcriptWidths(
    start(exonGRangesRestrict),
    end(exonGRangesRestrict))

absShiftVal <- abs(shiftValue)

ixSmallTransc <- which(transcWidth <= absShiftVal)

if(length(ixSmallTransc) > 0) {
  transcBig <- exonGRangesRestrict[-ixSmallTransc]
  cdsPosTRanscBig <- cdsPosTransc[-ixSmallTransc]
}
else {
  transcBig <- exonGRangesRestrict
}

overlapReads <-
  suppressWarnings(
    findOverlaps(originalAln, transcBig))

startOverlapReads <-
  split(start(
    alnGRanges[queryHits(overlapReads)]),
    factor(subjectHits(overlapReads)))

overlapReadsRle <-
  sapply(startOverlapReads, S4Vectors::Rle)

transcWithReads <-
  transcBig[as.numeric(names(overlapReadsRle))]

cdsPosTranscWithReads <-
  cdsPosTransc[as.numeric(names(overlapReadsRle))]

cdslengthwithreads <-
  lapply(seq_len(NROW(cdsPosTranscWithReads)),
         function(ixTransc) {
           cdsPosTranscWithReads[[ixTransc]][2] -
             cdsPosTranscWithReads[[ixTransc]][1] + 1
         })

newTranscWidth <-
```

```
  GenomicFeatures::transcriptWidths(
    start(transcWithReads),
    end(transcWithReads))

cdsPosTranscShifted <- do.call(rbind,
                               cdsPosTranscWithReads) +
  shiftValue

listeRangesCDS <-
  lapply(seq_len(NROW(cdsPosTranscShifted)),

         function(ixTransc) {
           max(1,
               cdsPosTranscShifted[ixTransc, 1]):
             cdsPosTranscShifted[ixTransc, 2]
         })

listeRanges5UTR <-
  lapply(seq_len(NROW(cdsPosTranscShifted)),
         function(ixTransc) {
           if((cdsPosTranscShifted[ixTransc,
                                   1] - 1) < 1) {
             0
           }
           else {
             max(1, shiftValue):(cdsPosTranscShifted[
               ixTransc,1] - 1)
           }
         })

listeRanges3UTR <-
  lapply(seq_len(NROW(cdsPosTranscShifted)),
         function(ixTransc) {
           (cdsPosTranscShifted[ixTransc, 2] +
              1):min(newTranscWidth[ixTransc],
                     newTranscWidth[ixTransc] +
                       shiftValue)
         })

binTransc <- applyShiftFeature(transcWithReads, 0)

strandInfo <-
```

```
    S4Vectors::runValue(strand(transcWithReads))

  shiftedTranscMatches <-
    lapply(seq_len(length(binTransc)), function(ixTransc) {

      if(strandInfo[[ixTransc]] == "-") {
        binTranscVal <- sort(binTransc[[ixTransc]],
                             decreasing = TRUE)
        txtail <- tail(binTranscVal, n = 1)
        binTranscVal <- c(binTranscVal,
                          seq(txtail-1, txtail-25))
      }
      else {
        binTranscVal <- sort(binTransc[[ixTransc]])
        txtail <- tail(binTranscVal, n = 1)
        binTranscVal <-
          c(binTranscVal, seq(txtail+1, txtail+25))
      }

      matchedReadsTransc <-
        match(sort(overlapReadsRle[[ixTransc]]),
              binTranscVal)

      matchedReadsCDS <-
        naTozeroRle(match(matchedReadsTransc,
                          listeRangesCDS[[ixTransc]]))

      matchedReads5UTR <-
        naTozeroRle(match(matchedReadsTransc,
                          listeRanges5UTR[[ixTransc]]))

      matchedReads3UTR <-
        naTozeroRle(match(matchedReadsTransc,
                          listeRanges3UTR[[ixTransc]]))

      if(length(matchedReadsCDS) > 0) {
        allCodonCounts <-
          aggregate(S4Vectors::runLength(
            matchedReadsCDS),
            by = list(ceiling(
              S4Vectors::runValue(matchedReadsCDS)/3)),
            FUN = sum)
```

```
    if(motifSize <= 3) {
      myCodonCounts <- allCodonCounts
    }
    else {
      if(motifSize == 6) {
        myCodonCounts <-
          allCodonCounts[1:(nrow(
            allCodonCounts) - 1), ]
      }
      else {
        if(motifSize == 9) {
          myCodonCounts <- allCodonCounts[2:(
            nrow(allCodonCounts) - 1), ]
        }
      }
    }
  }
  else {
    nbrCodons <-
      ceiling(length(listeRangesCDS[[ixTransc]])/
                motifSize)
    myCodonCounts <-
      data.frame(cbind(1:nbrCodons,
                       rep(0, nbrCodons)))
  }

  nbrCodons <-
    ceiling(cdslengthwithreads[[ixTransc]]/motifSize)

  myCodonCounts2 <-
    data.frame(cbind(1:nbrCodons, rep(0, nbrCodons)))

  names(myCodonCounts) <- c("codonID", "nbrReads")
  names(myCodonCounts2) <- c("codonID", "nbrReads2")

  myCodonCounts3 <-
    merge.data.frame(myCodonCounts,
                     myCodonCounts2,
                     by = 'codonID',
                     all = T)[,-3]
```

```
    names(myCodonCounts3) <- c("codonID", "nbrReads")

    list(c(sum(S4Vectors::runLength(matchedReadsCDS)),
           sum(S4Vectors::runLength(matchedReads5UTR)),
           sum(S4Vectors::runLength(matchedReads3UTR))),
         myCodonCounts3)
  })

names(shiftedTranscMatches) <- names(transcWithReads)

countsFeatures <-
  do.call(rbind,
          lapply(shiftedTranscMatches, `[[`, 1))

colnames(countsFeatures) <-
  c("CDS_counts", "fiveUTR_counts", "threeUTR_counts")
rownames(countsFeatures) <- names(shiftedTranscMatches)

chrInfo <-
  S4Vectors::runValue(
    GenomeInfoDb::seqnames(transcWithReads))
startInfo <- min(start(transcWithReads))
endInfo <- max(end(transcWithReads))
cdsInfo <- do.call(rbind, cdsPosTranscWithReads)
cdsLength <- cdsInfo[, 2] - cdsInfo[, 1] + 1
cdsStart <- cdsInfo[, 1]
cdsEnd <- cdsInfo[, 2]

countsData <-
  cbind(as.character(rownames(countsFeatures)),
        as.character(unlist(chrInfo)),
        as.character(unlist(strandInfo)),
        startInfo, endInfo, newTranscWidth, cdsStart,
        cdsEnd, cdsLength, countsFeatures)

colnames(countsData) <-
  c("gene", "chr", "strand", "transc_genomic_start",
    "transc_genomic_end", "transc_length", "orf_start",
    "orf_end", "orf_length", colnames(countsFeatures))

codonReadCoverage <-
  lapply(shiftedTranscMatches, `[[`, 2)
```

```
  names(codonReadCoverage) <- names(shiftedTranscMatches)

  return(list(as.data.frame(countsData),
              codonReadCoverage))

}

plotSummarizedCov <- function (covSummarized)
{
  if (!inherits(covSummarized, "list")) {
    stop("The covSummarized object is not a list!\n")
  }
  else {
    if (!is(covSummarized[[1]], "GRanges")) {
      stop("The covSummarized object is not a list of
           GRanges objects!\n")
    }
  }
  listPlotSum <- lapply(covSummarized, function(iSumCov) {
    maxPeak <- max(iSumCov$values)
    maxPeakPos <- start(iSumCov)[which(iSumCov$values ==
                                         maxPeak)]
    if (maxPeak <= 100) {
      yLab <- "% of reads"
    }
    else {
      yLab <- "Number of Reads"
    }
    iPlot <- ggplot(iSumCov, ggplot2::aes(start, values)) +
      geom_line(color = black) + geom_point(color = black) +
      xlim(0, 30) +
      labs(x = "Distance from start codon (nt)",
           y = yLab,
           title = "Calculating p-site offset") +
      scale_y_continuous(labels = comma_format()) +
      paper_theme
    return(iPlot)
  })

  return(listPlotSum)
}
```

```
load_gff <- function() {
  cat("\nPlease provide directory to GFF file\n")
  annotation.gff <- readline("Choice: ")

  txdb <- GenomicFeatures::makeTxDbFromGFF(
    annotation.gff, format = "gff3", circ_seqs = character())
  txids <- AnnotationDbi::keys(txdb, keytype="TXNAME")
  cdsTransc <-
    GenomicFeatures::cdsBy(txdb, by = "tx", use.names = T)
  exonGRanges <-
    GenomicFeatures::exonsBy(txdb, by = "tx", use.names = T)
  cdsPosTransc <-
    RiboProfiling::orfRelativePos(cdsTransc, exonGRanges)

  tx <- list(
    "txdb" = txdb,
    "txids" = txids,
    "cdsTransc" = cdsTransc,
    "exonGRanges" = exonGRanges,
    "cdsPosTransc" = cdsPosTransc)

  assign("tx", tx, envir = .GlobalEnv)
}
```

Quicker wrangling

```
# Necessary for computations of multiple samples, uses
# non-standard evaluation

obey <- function(string) {
  eval(parse(text = string), envir = .GlobalEnv)
}

# Calculating standard deviation from mean for data curves
# Used in to bin genes for protein length vs. enrichment
# and to calculate read count thresholds

quant_sd <-
  function(data, predictor, response, quantiles = 5) {
```

```
  dt <- deparse(substitute(data))

  pred <- deparse(substitute(predictor))
  pred.dt <- paste0(dt, "$", pred)
  cat.sd <- paste0(pred, ".sd")
  obey(paste0(
    "pred.range <- as.character(floor(quantile(", pred.dt,
    ", probs = seq(0, 1, ", as.character(1/quantiles),
    "))))"
  ))

  response <- deparse(substitute(response))

  for (i in 1:(length(pred.range) - 1)) {

    obey(paste0(
      "pred.mean <- mean(", dt, "[", pred, " >= ",
      pred.range[[i]], " & ", pred, " <= ",
      pred.range[[i + 1]], "]$", response, ")"
    ))
    obey(paste0(
      "pred.sd <- sd(", dt, "[", pred, " >= ",
      pred.range[[i]], " & ", pred, " <= ",
      pred.range[[i + 1]], "]$", response, ")"
    ))
    obey(paste0(
      dt, "[", pred, " >= ", pred.range[[i]],
      " & ", pred, " <= ", pred.range[[i + 1]], " & ",
      response, " >= ", pred.mean, " - 3*", pred.sd, " & ",
      response, " <= ", pred.mean, " + 3*", pred.sd, ", ",
      cat.sd, " := 3]"
    ))
    obey(paste0(
      dt, "[", pred, " >= ", pred.range[[i]],
      " & ", pred, " <= ", pred.range[[i + 1]], " & ",
      response, " >= ", pred.mean, " - 2*", pred.sd, " & ",
      response, " <= ", pred.mean, " + 2*", pred.sd, ", ",
      cat.sd, " := 2]"
    ))
    obey(paste0(
      dt, "[", pred, " >= ", pred.range[[i]],
      " & ", pred, " <= ", pred.range[[i + 1]], " & ",
```

```
        response, " >= ", pred.mean, " - ", pred.sd, " & ",
        response, " <= ", pred.mean, " + ", pred.sd, ", ",
        cat.sd, " := 1]"
      ))
      obey(paste0(
        dt, "[", pred, " > ", pred.range[[i]], " & ",
        pred, " <= ", pred.range[[i + 1]], " & ",
        "is.na(", cat.sd, "), ",
        cat.sd, " := 4]"
      ))
    }
  }

# Convert zero values to NA

zero_na <- function(DT, list = names(DT)) {
  invisible(lapply(
    list, function(.name)
      set(DT, which(DT[[.name]] == 0),
          j = .name, value = NA)))
}

# Convert specified value to NA

val_na <- function(DT, var,  list = names(DT)) {
  invisible(lapply(
    list, function(.name)
      set(DT, which(DT[[.name]] == var), j = .name, var = NA)
  ))
}

# Convert NA value to zero

na_zero <- function(DT, list = names(DT)) {
  invisible(lapply(
    list,function(.name)
      set(DT, which(is.na(DT[[.name]])),
          j = .name,value = 0)))
}

# Convert NaN value to NA
```

```
nan_na <- function(DT, list = names(DT)) {
  invisible(lapply(
    list,function(.name)
      set(DT, which(is.nan(DT[[.name]])),
          j = .name,value = NA)))
}

# Convert NaN value to zero

nan_zero <- function(DT, list = names(DT)) {
  invisible(lapply(
    list,function(.name)
      set(DT, which(is.nan(DT[[.name]])),
          j = .name,value = 0)))
}

# Convert infinite values to NA, necessary when comparing
# enrichment scores where membrane/cytosol ribosomes have
# zero reads

inf_na <- function(DT, list = names(DT)) {
  invisible(lapply(
    list,function(.name)
      set(DT, which(is.infinite(DT[[.name]])),
          j = .name,value = NA)))
}

# Calculate enrichment scores between membrane/cytosol
# samples

calc_enrich <- function(df, output, mem, sol) {
  output <- enquo(output)
  mem <- enquo(mem)
  sol <- enquo(sol)

  es_name <- quo_name(output)

  df %>% mutate(!! es_name := log2((((!! mem) / 1e6) /
                                    ((!! sol) / 1e6)))
}
```

```
# Calculate nascent chains and ribosome usage

calc_expression <- function(df, output, input,
                            norm = NULL) {
  output <- enquo(output)
  input <- enquo(input)
  norm <- enquo(norm)

  output <- quo_name(output)

  if (grepl("tpm", output)) {
    df %>%
      mutate(rpk = (!! input)/((!! norm)/1e3)) %>%
      ungroup() %>%
      mutate(!! output := rpk/(sum(rpk)/1e6)) %>%
      select(-rpk)
  } else if (grepl("rpm", output)) {
    df %>%
      group_by(.id) %>%
      mutate(num = sum(smoothed)) %>%
      group_by(mask, add = TRUE) %>%
      mutate(den = sum(smoothed),
             den = first(den)) %>%
      ungroup() %>%
      mutate(rScaleFactor = num/den,
             rpm = nbrReads * rScaleFactor) %>%
      filter(mask == FALSE) %>%
      mutate(rpm = rpm/(sum(rpm, na.rm = TRUE)/1e6)) %>%
      group_by(.id) %>%
      mutate(rpm = sum(rpm, na.rm = TRUE)) %>%
      select(-num, -den, -rScaleFactor)
  }
}

load_files <- function() {

  cat("\nHow many files would you like to analyze?\n")

  num.files <- suppressWarnings(as.integer(readline()))

  if (is.na(num.files)) {
```

```
    cat("\nPlease provide answer as an integer!\n\n")
  } else {
    files <- c()
    sample_names <- c()
    for (i in 1:num.files) {
      cat(paste0(
        "\nPlease input directory for file # ", i, "\n"))
      files[i] <- readline("Choice: ")

      cat("\nWhat sample is this?\n")
      sample_names[i] <- readline("Choice: ")
    }
    cat("\n")

    assign("files", files, envir = .GlobalEnv)
    assign("sample_names", sample_names, envir = .GlobalEnv)
  }
}
```

Mask generator

```
# Necessary functions
strpick <- function(string, split = " ", pos = NA) {
  nStr <-
    length(unlist(strsplit(string, split, fixed = TRUE)))
  str <- unlist(strsplit(string, split, fixed = TRUE))

  if (is.na(pos)) {
    return(str)
  }
  else {
    if (nStr == 1) {
      return(str[[nStr]])
    }
    else if (pos > nStr) {
      return(str[[nStr]])
    }
    else {
      return(str[[pos]])
```

```
    }
  }
}

partition_sequence <- function(string, footprint = 28) {
  gene.l <-
    length(unlist(strsplit(string, "", fixed = TRUE)))
  gene.seq <- unlist(strsplit(string, "", fixed = TRUE))

  if (gene.l <= footprint) {
    stop(paste0(
      "Gene length too low to partition for ribosome
      profiling (",
      footprint, " nt footprints)!\n")
    )
  }
  else {
    txt.fasta <- ""
    for (i in 1:(gene.l - footprint)) {
      start <- i
      start
      end <- i + (footprint - 1)
      end
      txt.sample <-
        paste0(gene.seq[start:end], collapse = "")
      if (i == 1) {
        txt.fasta <- txt.sample
      }
      else {
        txt.fasta <- paste0(txt.fasta, "\n", txt.sample)
      }
    }
    cmd <- paste0(
      "ORF scrambled... ", i, " unique fragments ",
      footprint, " nucleotides long produced. \n"
    )
    cat(cmd)
    return(txt.fasta)
  }
}

# Read CDS fasta file using Biostrings
```

```
annotation_cds <-
  "cds-fasta-path"

dna_fasta <- Biostrings::readDNAStringSet(annotation_cds)
seq.names <- names(dna_fasta)
gene.seq <- paste(dna_fasta)

# Make table of genes and their sequences
dna_dt <- data.table(seq.names, gene.seq)

# Get gene id (this changes depending on the way that
# Biostrings reads names)
dna_dt[, id := sapply(seq.names,
                      function(x) strpick(x, split = " ",
                                          pos = 1))]

# Get gene name (not all files will produce a gene name)
dna_dt[, gene.name := sapply(seq.names,
                             function(x) strpick(x,
                                                 split = " ",
                                                 pos = 2))]

# Tidy up data table
dna_dt[, seq.names := NULL]
setcolorder(dna_dt, c('id', 'gene.name', 'gene.seq'))

# Partition DNA sequences into 28mer fragments and generate
# DNA fasta filerepresenting pseudo Ribo-seq data. Use
# this data to run Hisat2 on to get sort.bam files needed
# to create mask
dna_dt[,
       gene.seq.mix := lapply(
         gene.seq,
         function(x) partition_sequence(x))]

output_txt <- "kphaffii_mix.txt"

write.table(dna_dt$gene.seq.mix,
            file = output_txt,
            quote = FALSE,
            row.names = FALSE,
            col.names = FALSE,
```

```
            sep = '\n')

output_sam <- "kphaffii_mix.sam"

annotation.hisat <-
  "genomic-fasta-path"

command <- paste0(
  "/Users/Shared/Repository/miniconda3/envs/seq/bin/hisat2 ",
  "-x ", annotation.hisat, " ",
  "-U ", file.txt, " ",
  "-p 7 ",
  "-r ",
  "-S ", output.sam
)
system(command)
```

Protein sequence predictions

```
# Analysis of protein sequence features from protein fasta
# files
annotation.protein <- "protein-fasta-path"

protein.fasta <-
    Biostrings::readAAStringSet(annotation.protein)
  seq.names <- names(protein.fasta)
  protein.seq <- paste(protein.fasta)
  rm(protein.fasta)
  protein.dt <- data.table(seq.names, protein.seq)
  protein.dt[, id := as.character(lapply(
    seq.names, function(x) strsplit(x, " ")[[1]][[1]]
  ))]
  setkey(protein.dt, 'id')
  protein.dt[, protein.l := nchar(protein.seq)]
  protein.dt <- protein.dt[, .(id, protein.l, protein.seq)]

# Generating best match homology from blastP outputs
blast.file <- "blastP-output-path"
```

```
blast.dt <- as.data.table(read.delim(
  file = blast.file,
  header = FALSE,
  col.names = c("id", "match.id", "aln.percentage",
                "aln.length", "mismatches", "gapOpenings",
                "query.start", "query.end", "match.start",
                "match.end", "evalue", "bitscore")
))
setkey(blast.dt, id)

unique.ids <- unique(blast.dt$id)

for (i in 1:length(unique.ids)) {
  cat(paste0("Generating best match for ",
             unique.ids[[i]], "...\n"))

  gene <- blast.dt[id == unique.ids[[i]]]

  tmp <- gene[evalue == min(gene$evalue)]
  if (nrow(tmp) > 1) {tmp <-
    tmp[bitscore == max(tmp$bitscore)]}
  if (nrow(tmp) > 1) {tmp <- tmp[1]}

  if (i == 1) {tmp.dt <- tmp} else {
    tmp.dt <- rbind(tmp.dt, tmp)}
}

blast.dt <- tmp.dt[,.(id, match.id)]

# Signal peptide predictions from SignalP 5.0

signalp.dt <- as.data.table(
    read.delim2(
      file = "signalp5.0-output-path",
      col.names = c("id", "prediction", "sp.perc",
                    "other.perc", "position"),
      as.is = c(1:5), header = FALSE
    )) %>%
    slice(3:nrow(signalp.dt)) %>%
    mutate(sp.sp = if_else(
      prediction == "SP(Sec/SPI)", TRUE, FALSE),
           position = na_if(position, "")) %>%
```

```
    mutate(sp.l = str_sub(position, 12, 13)) %>%
    select(-prediction, -other.perc) %>%
    rename("sp.position"="position")

# GPI predictions from GPI pred
gpipred.dt <- as.data.table(
    read.delim2(file = "gpipred-output-path",
                col.names = c("id", "gpi.fpr",
                              "gpi.omega"),
                as.is = c(1:3),
                header = FALSE))

  gpipred.dt[, id := gsub(">", "", id)]
  gpipred.dt[, gpi.fpr := gsub("FPrate:", "", gpi.fpr)]
  gpipred.dt[, gpi.omega := gsub("OMEGA:", "", gpi.omega)]
  gpipred.dt[, gpi.aa := sapply(
    gpi.omega,
    function(x) unlist(strsplit(x, "-",
                                fixed = TRUE))[[1]])]
  gpipred.dt[, gpi.index := sapply(
    gpi.omega,
    function(x) unlist(strsplit(x, "-",
                                fixed = TRUE))[[2]])]
  gpipred.dt[, gpi.omega := NULL]
  gpipred.dt[,
             gpi.specificity.index :=
               (1 - as.numeric(gpi.fpr)) * 100]
  gpipred.dt[, gpi.prediction := 0]
  gpipred.dt[gpi.specificity.index >= 99.0,
             gpi.prediction := 1]
  gpipred.dt[gpi.specificity.index >= 99.5,
             gpi.prediction := 2]
  gpipred.dt[gpi.specificity.index >= 99.9,
             gpi.prediction := 3]

# Subcellular localization predictions from DeepLoc
file_name <- "deeploc-output-path"

deeploc.dt <- read_csv(file = file_name,
                       col_names = TRUE) %>%
  rename(id = `Entry ID`, dl.loc = "Localization",
         dl.type = "Type") %>%
```

```
  select(id, dl.loc, dl.type)

# Signal peptide and transmembrane domain predictions from
# TopCons

annotation.TopCons <- "topcons-output-path"

if (exists("annotation.TopCons")) {
  sys.cmd <- paste0(
    "awk '{print $7,$4,$3}' ", annotation.TopCons
  )
  topcons.dt <- as.data.table(system(sys.cmd, intern = TRUE))
  topcons.dt[, id := as.character(lapply(
    V1, function(x) strsplit(x, " ")[[1]][1]))]
  topcons.dt[, tc.sp := as.logical(lapply(
    V1, function(x) strsplit(x, " ")[[1]][2]))]
  topcons.dt[, tc.tmd := as.integer(lapply(
    V1, function(x) strsplit(x, " ")[[1]][3]))]
  topcons.dt[, V1 := NULL]
  setkey(topcons.dt, "id")
}

# Transmembrane domain predictions from TMHMM

annotation.TMHMM <- "TMHMM-output-path"

if (exists("annotation.TMHMM")) {
  prob.cmd <- paste0(
    "grep \"prob\" ", annotation.TMHMM
  )
  probs <- as.data.table(system(prob.cmd, intern = TRUE))
  probs[, id := as.character(lapply(
    V1, function(x) strsplit(x, " ")[[1]][2]))]
  probs[, tmd.tmhmm.p := as.numeric(lapply(
    V1, function(x) strsplit(x, " ")[[1]][14]))]
  setkey(probs, "id")

  pred.cmd <- paste0(
    "grep \"predicted\" ", annotation.TMHMM
  )
  preds <- as.data.table(system(pred.cmd, intern = TRUE))
  preds[, id := as.character(lapply(
```

```
    V1, function(x) strsplit(x, " ")[[1]][2]))]
  preds[, tmd.tmhmm.n := as.numeric(lapply(
    V1, function(x) strsplit(x, " ")[[1]][8]))]
  setkey(preds, "id")

  tmhmm.dt <- merge(preds[,
                          .(id, tmd.tmhmm.n)],
                    probs[, .(id, tmd.tmhmm.p)],
                    all = TRUE)
  rm(preds, probs)
}

# Ontological predictions from EggNOG 2.0
egg_file <- "eggnog2.0-output-path"

if (TRUE) {
  ## Read eggnog output file into R as a data table, setkey
  ## to 'id'
  eggnog.dt <- as.data.table(read.delim2(
    file =  egg_file,
    header = FALSE,
    na.strings = c("NA|NA|NA", ""),
    col.names = c("id", "egg.ortholog", "egg.value",
                  "egg.score", "egg.tax",
                  "eggnog.name",    "egg.GO",
                  "egg.EC", "kegg.ko", "kegg.pathway",
                  "kegg.module",    "kegg.rxn",
                  "kegg.rclass","egg.brite", "kegg.tc",
                  "egg.CAZy", "BiGG.rxn", "egg.annotlvl",
                  "egg.og","egg.bestog", "eggnog.cog",
                  "description.eggnog")
  ))
  eggnog.dt <- eggnog.dt[,.(id, eggnog.name, eggnog.cog,
                            description.eggnog)]
  setkey(eggnog.dt, 'id')

  ## Convert cog score output into list of cog scores
  eggnog.dt[, eggnog.cog := as.character(eggnog.cog)]
  eggnog.dt[is.na(eggnog.cog), eggnog.cog := "S"]
  eggnog.dt[, eggnog.cog := strsplit(eggnog.cog, '')]
```

```
## Create NEW data table by duplicating id n times,
## where n = length(cog.list), and pairing with unique
## cog score from cog.list
for (i in 1:nrow(eggnog.dt)) {

  dt <- as.data.table(eggnog.dt$id[[i]])

  for (j in 1:length(unlist(eggnog.dt[dt]$eggnog.cog))) {
    eggnog.cog <- unlist(eggnog.dt[dt]$eggnog.cog)[j]
    tmp <- copy(dt)
    tmp[, V2 := eggnog.cog]
    if (j == 1) {
      dt.hold <- tmp
    } else {
      dt.hold <- merge(dt.hold, tmp, all = TRUE)
    }
  }
  if (i == 1) {
    dt.final <- dt.hold
  } else {
    dt.final <- merge(dt.final, dt.hold, all = TRUE)
  }
}
cog.dt <- dt.final; rm(dt.final)
names(cog.dt) <- c('id', 'eggnog.cog')

## Create cog.table with relative cog frequency for
## hierarchichal clustering
cog.table <- as.data.table(table(cog.dt$eggnog.cog))
names(cog.table) <- c('eggnog.cog', 'cog.n')
cog.table[, cog.freq := cog.n/(sum(cog.table$cog.n))]
setkey(cog.table, "cog.freq")
egg.hierarchy <- rev(cog.table$eggnog.cog)

## Cluster duplicated cog scores into highest frequency
## cog from cog.table
while (nrow(cog.dt[duplicated(cog.dt$id)]) > 0) {
  dt <- cog.dt[duplicated(cog.dt$id)]

  id <- dt$id[[1]]
```

```
  obey(paste0("match.list <- dt[, id == '", id, "']"))
  potentials <- dt[match.list]$eggnog.cog

  tmp.best <- length(egg.hierarchy)
  for (i in 1:length(potentials)) {
    place <- lapply(egg.hierarchy,
                    function(x) grep(potentials[[i]], x))
    place <- which(place == 1)
    if (place < tmp.best) {
      tmp.best  <- place
    }
  }
  best.fit <- egg.hierarchy[[tmp.best]]
  obey(paste0("cog.dt[id == '", id,
              "', eggnog.cog := '", best.fit, "']"))
  cog.dt <- unique(cog.dt)
  cat(paste0(
    "\nCHANGE: ", id,
    "'s COG value will be changed to ", best.fit, "\n"))
  cat(paste0(
    "Duplicates left: ",
    as.character(nrow(cog.dt[duplicated(cog.dt$id)])),
    "\n"))

}

eggnog.dt <- eggnog.dt %>% slice(-(1:3))
cog.dt <- cog.dt %>% slice(-(1:3))

eggnog.dt <- merge(
  eggnog.dt[,.(id, eggnog.name, description.eggnog)],
  cog.dt,
  all = TRUE)

## Get COG tranlsations
cog.def <- as.data.table(read.delim2(
  file = "def-COG-path",
  header = TRUE,
  as.is = (c(1:4)),
  col.names = c("eggnog.cog", "category", "subcategory",
                "color")
))
```

```
  eggnog.dt <- merge(eggnog.dt, cog.def,
                     by.x = "eggnog.cog",
                     by.y = "eggnog.cog",
                     all = TRUE)
  setcolorder(eggnog.dt, c(2:4,1,5:7))
  setkey(eggnog.dt, id)
  eggnog.dt <- eggnog.dt[!is.na(id)]
  eggnog.dt[, color := sapply(color,
                              function(x) rand_color(x))]

  rm(cog.def, cog.dt, cog.table, dt, dt.hold, tmp)
}
```

Ribo-Seq pipeline

```
# Load files and annotation file that will be used for
# analysis
load_files()
load_gff()

# Convert reference to BAM files.
files <- lapply(files, Rsamtools::BamFile)

# Convert aligned reads from BAM file into GAlignments
# object.
aln <- lapply(files, GenomicAlignments::readGAlignments)

# Convert GAlignments object to End (3') positions.
alnGRanges <-
  lapply(aln, RiboProfiling::readsToStartOrEnd,
         what = "end")

# Returns a GRanges object containing the flank size around
# the transcriptional
# start site (TSS) for selected coding sequence (CDS).
flank_size <- 30
oneBinRanges <- lapply(seq(length(files)),
                       function(x)
```

```
                          RiboProfiling::aroundPromoter(
                            txdb = tx$txdb,
                            alnGRanges = alnGRanges[[x]],
                            percBestExpressed = 0.01,
                            flankSize = flank_size))

# Create histogram of match length distribution of reads.
matchLenDistr <- lapply(aln, RiboProfiling::histMatchLength)

# Create vector of match lengths with read counts greater
# than 3000.
match_lengths <- lapply(seq(length(files)),
                        function(x)
                          as.numeric(as.character(unlist(
                            matchLenDistr[[x]][[1]]$
                              matchSize)))[
                              which(matchLenDistr[[x]][[1]]$
                                      counts > 3000)])

# Calculate summarized read coverages around TSS for
# specified match lengths.
listPromoterCov <-
  lapply(seq(length(files)),
         function(x) RiboProfiling::readStartCov(
           alnGRanges = alnGRanges[[x]],
           oneBinRanges = oneBinRanges[[x]],
           matchSize = match_lengths[[x]],
           fixedInterval = c(-flank_size,
                             flank_size),
           renameChr = "aroundTSS",
           charPerc = "sum"))

# Calculate psite offsets from listPromoterCov.
shift <- sapply(seq(length(files)),
                function(x) as.numeric(
                  listPromoterCov[[x]]$sumUp@ranges@start[
                    which.max(listPromoterCov[[x]]$
                                sumUp@elementMetadata@
                                listData$values)]))

# Applies psite offset on read start along trascript and
# returns the following:
```

```
# 1. Information on ORF including names, position, lengths,
# and counts on 5'UTR, CDS, and 3'UTR after offset is
# applied.
# 2. List of dataframes for each ORF containing read counts
# per codon.
counts <- lapply(seq(length(files)),
                 function(x) countReads(
                   exonGRanges =
                     tx$exonGRanges[names(tx$cdsPosTransc)],
                   cdsPosTransc = tx$cdsPosTransc,
                   alnGRanges = alnGRanges[[x]],
                   originalAln = aln[[x]],
                   shiftValue = shift[[x]],
                   motifSize = 3))

# Collapses list of dataframes for each ORF containing read
# counts per codon
# into one dataframe.
counts <- lapply(seq(length(files)),
                 function(x) ldply(counts[[x]][[2]]))

# Apply masking to read count dataframe.
masked <- lapply(seq(length(files)), function(x)
  full_join(mask, counts[[x]], by = c(".id", "codonID")) %>%
    filter(codonID > 5 & codonID <= theory.l - 5) %>%
    replace_na(list(nbrReads = 0)) %>%
    select(-single.reads, -multiple.reads, -diff, -theory.l))

# Calculate average reads per codon for first 100 codons per
# gene to reduce biases associated with increased read counts
# at beginning of transcripts.
rpc <- lapply(seq(length(files)), function(x)
  masked[[x]] %>%
    filter(codonID <= 100 & mask == FALSE) %>%
    group_by(.id) %>%
    mutate(rpc.100 = mean(nbrReads),
           rpc.100 = na_if(rpc.100, 0)))

# Normalize reads per codon by average reads per codon for
# first 100 codons and determine reads per gene to establish
# cut off criterion for metagene analysis
norm <- lapply(seq(length(files)), function(x)
```

```
  full_join(masked[[x]], rpc[[x]]) %>%
    filter(mask == FALSE) %>%
    group_by(.id) %>%
    fill(rpc.100) %>%
    replace_na(list(rpc.100 = 0)) %>%
    mutate(rpc.100 = na_if(rpc.100, 0),
           norm.100 = nbrReads/rpc.100,
           rpg = mean(nbrReads)) %>%
    replace_na(list(norm.100 = 0)))

# Perform metagene analysis using non-enriched genes with
# adequate reads per gene and average reads per codon for
# first 100 codons. Metagene analysis calculates average
# normalized reads per codon for all genes. This average
# normalized read per codon is smoothed using a rolling mean
# and rolling median.
meta.counts <- lapply(seq(length(files)), function(x)
  norm[[x]] %>%
    filter(.id %in% not_enriched & rpg > 0.5 &
             rpc.100 > 0.5) %>%
    group_by(codonID) %>%
    summarise(counts = sum(norm.100 > 0)))


meta <- lapply(seq(length(files)), function(x)
  full_join(norm[[x]], meta.counts[[x]]) %>%
    filter(.id %in% not_enriched & rpg > 0.5 &
             rpc.100 > 0.5) %>%
    group_by(codonID) %>%
    mutate(meta = if_else(counts >= 2,
                          mean(norm.100,
                               trim = 0.05,
                               na.rm = TRUE),
                          median(norm.100,
                                 trim = 0.05,
                                 na.rm = TRUE))) %>%
    select(codonID, meta) %>%
    distinct() %>%
    ungroup() %>%
    arrange(codonID))
```

```
meta.dt <- lapply(seq(length(files)), function(x)
  meta[[x]] %>%
    full_join(meta[[x]] %>%
                filter(codonID >= 0 & codonID <= 100) %>%
                mutate(
                  smoothed = rollapply(meta,
                                       width = 10,
                                       FUN = mean,
                                       partial = TRUE))) %>%
    full_join(meta[[x]] %>%
                filter(codonID > 100 &
                         codonID <= 500) %>%
                mutate(smoothed = rollapply(
                  meta,
                  width = 100,
                  FUN = mean,
                  partial = TRUE))) %>%
    full_join(meta[[x]] %>%
                filter(codonID > codon_limit[[x]]) %>%
                mutate(smoothed = rollapply(
                  meta,
                  width = 1000,
                  FUN = median,
                  partial = TRUE))) %>%
    filter(!is.na(smoothed)) %>%
    mutate(smoothed = na_if(smoothed, 0)) %>%
    fill(smoothed))

# Scale number of reads per codon per gene after normalizing
# reads per codon per gene by smoothed metagene reads per
# codon. Use scaled reads per codon per gene to calculate
# expression in RPM and cTPM.
scaled <- lapply(seq(length(files)), function(x)
  full_join(masked[[x]], meta.dt[[x]]) %>%
    group_by(`.id`) %>%
    mutate(num = sum(smoothed, na.rm = TRUE)) %>%
    group_by(mask, add = TRUE) %>%
    mutate(den = sum(smoothed, na.rm = TRUE)) %>%
    filter(mask == FALSE) %>%
    mutate(rScaleFactor = num/den,
           crpm = nbrReads * rScaleFactor) %>%
```

```
    replace_na(list(crpm = 0)) %>%
    ungroup() %>%
    mutate(crpm = crpm/sum(crpm)*1e6,
           scaled = if_else(codonID <= 500,
                            nbrReads/smoothed, nbrReads)) %>%
    mutate(scaled = if_else(is.na(scaled), 0, scaled)) %>%
    group_by(`.id`) %>%
    mutate(crpm = sum(crpm),
           reads = sum(nbrReads),
           scaled = sum(scaled)) %>%
    select(-den, -num, -rScaleFactor, -codonID, -meta,
           -smoothed, -nbrReads, -mask) %>%
    distinct() %>%
    calc_expression(output = ctpm,
                    input = scaled,
                    norm = pseudo.l) %>%
    select(-pseudo.l, -scaled) %>%
    mutate(rpm = reads/sum(reads)*1e6) %>%
    relocate(.id, reads, rpm, crpm, ctpm) %>%
    rename(id = .id,
           !! paste0(sample_names[[x]], ".reads") := reads,
           !! paste0(sample_names[[x]], ".rpm") := rpm,
           !! paste0(sample_names[[x]], ".crpm") := crpm,
           !! paste0(sample_names[[x]], ".ctpm") := ctpm))

# Collapse different samples' dataframes into one dataframe
# and calculate membrane enrichment.
dt <- Reduce(function(x, y) full_join(x, y), scaled) %>%
  mutate(id = if_else(id == "HSA-T1", "HSA", id))
```

Python processing

Sucrose gradient analysis

```
# Visualizing sucrose fractionation bands to determine
# ideal RNAse concentration for nuclease digestion
import pandas as pd
```

```
import matplotlib.pyplot as plt
from scipy import integrate
from directory_modifications import File


class Fraction(File):
    def __init__(self, file_name):
        self.cell_line = file_name.split('/')[-2]
        File.__init__(self, file_name)

    def load_data(self):
        csv_file = self.show_full_name()
        df = pd.read_csv(csv_file,
                         skiprows=range(0, 42),
                         usecols=lambda x:
                           x.strip() in ['Position(mm)',
                           'Absorbance'])
        return df

    def get_cell(self):
        return self.cell_line


def _range_for_max(data_list):

    min_val = 15
    max_val = 25

    data_range = []
    for data in data_list:
        tmp_range = [i for i,
        j in enumerate(data['x']) if min_val < j < max_val]
        data_range.append(tmp_range)

    return data_range


def integrate_peak(file_object):
    data_list = order_data(file_object)
    subset_peak = _range_for_max(data_list)
    data_peak = _subset_data(data_list, subset_peak)
```

```
    whole_integration = []
    for data in data_list:
        tmp_int = integrate.trapz(data['y'], data['x'])
        whole_integration.append(tmp_int)

    peak_integration = []
    for data in data_peak:
        tmp_int = integrate.trapz(data['y'], data['x'])
        peak_integration.append(tmp_int)

    print('Percentage of 80s peak absorbance:\n')
    for i in range(len(file_object)):
        percentage = peak_integration[i]/
        whole_integration[i] * 100
        print('%s: %.2f %%' % (file_object[i].get_name(),
        percentage))
    else:
        print()

    return


def order_data(file_object):

    neighbors = 15
    data_list = []
    for file in file_object:
        df = file.load_data()
        df_rolling_mean =
        df.rolling(neighbors).mean().fillna(0)
        headers = list(df_rolling_mean)
        x = df_rolling_mean[headers[0]]
        y = df_rolling_mean[headers[1]]
        data_dic = {'x': x, 'y': y}
        data_list.append(data_dic)

    return data_list


def single_plot_data(file_object):

    data_list = order_data(file_object)
```

```
    data_range = _range_for_max(data_list)
    (max_places, max_val) = _find_max_indices(data_list,
    data_range)

    integrate_peak(file_object)

    plt.plot(data_list[0]['x'], data_list[0]['y'],
             label=file_object[0].get_cell() + ' ' +
             file_object[0].get_name(),
             linewidth=1,
             alpha=0.85)
    plt.xlim(10, 70)
    plt.ylim(min(data_list[0]['y']), max_val + 0.01)

    plt.xlabel('Position (mm)')
    plt.ylabel('Absorbance (280 um)')
    plt.title('Sucrose Fractionation Absorbance Readings')
    plt.legend(shadow=True)

    print('Would you like to display (d) or save (s) image?')
    choice = input('Choice: ').strip()
    print('\n')

    if choice == "d":
        plt.show()
    elif choice == "s":
        print('What would you like to name output file?')
        file_name = input('Choice: ').strip()
        print('\n')
        plt.savefig("UVspec/Sucrose_Fractions/images/" +
        file_name + ".svg", format="svg")

    return


def plot_data(file_object):

    # Load data from files. Store each file into x and y
    # values as a dictionary of x and y keys with lists
    # containing
    # data. Each file's data is read into one of these
    # dictionaries and stored in a list containing all of the
```

```
    # data.
    data_list = order_data(file_object)

    # Determine range of values to look for max
    data_range = _range_for_max(data_list)

    # Determine indices of max peak with data_range
    (max_places, max_val) = _find_max_indices(data_list,
    data_range)

    minimum_count = min(max_places)
    place_holder = max_places.index(minimum_count)

    data_shift = _shift_data(data_list, max_places,
    place_holder)

    integrate_peak(file_object)

    # Plot data
    minimum = min(data_list[0]['y'])

    plt.figure()
    for i in range(len(file_object)):
        plt.plot(data_list[i]['x'] - data_shift[i],
        data_list[i]['y'],
                 label=file_object[i].get_cell() + ' ' +
                 file_object[i].get_name(),
                 linewidth=1,
                 alpha=0.85)

        if min(data_list[i]['y']) < minimum:
            minimum = min(data_list[i]['y'])

    plt.xlim(10, 75)
    plt.ylim(minimum, max_val + max_val/10)

    plt.xlabel('Position (mm)')
    plt.ylabel('Absorbance (280 um)')
    plt.title('Sucrose Fractionation Absorbance Readings')
    plt.legend(shadow=True)

    print('Would you like to display (d) or save (s) image?')
```

```
    choice = input('Choice: ').strip()
    print('\n')

    if choice == "d":
        plt.show()
    elif choice == "s":
        print('What would you like to name output file?')
        file_name = input('Choice: ').strip()
        print('\n')
        plt.savefig("UVspec/Sucrose_Fractions/images/" +
        file_name + ".svg", format="svg")

    return


def _find_max_indices(data_list, data_range):
    max_places = []
    max_val = 0
    for index in range(len(data_list)):
        tmp_data = data_list[index]['y']
        tmp_data_range =
        tmp_data[data_range[index][0]: data_range[index][-1]]
        tmp_max = max(tmp_data_range)

        if tmp_max > max_val:
            max_val = tmp_max

        tmp_place = tmp_data[tmp_data == tmp_max].index[0]
        max_places.append(tmp_place)

    return max_places, max_val


def _shift_data(data_list, max_places, place_holder):
    data_shift = []
    for i in range(len(data_list)):
        tmp_shift = data_list[i]['x'][max_places[i]] -
        data_list[i]['x'][max_places[place_holder]]
        data_shift.append(tmp_shift)

    return data_shift
```

```
def _subset_data(data_list, subset):
    sub_data = []
    for i in range(len(data_list)):
        data_y = data_list[i]['y']
        data_x = data_list[i]['x']

        subset_y = data_y[subset[i][0]: subset[i][-1]]
        subset_x = data_x[subset[i][0]: subset[i][-1]]

        data_dic = {'x': subset_x, 'y': subset_y}
        sub_data.append(data_dic)

    return sub_data
```