

Title	Development of a ribosome profiling protocol to study translation in the yeast <i>Kluyveromyces marxianus</i>
Authors	Fenton, Darren
Publication date	2023-04-25
Original Citation	Fenton, D. 2022. Development of a ribosome profiling protocol to study translation in the yeast <i>Kluyveromyces marxianus</i> . PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2022, Darren Fenton. - <a href="https://creativecommons.org/licenses/by-nc-nd/4.0/">https://creativecommons.org/licenses/by-nc-nd/4.0/</a>
Download date	2024-04-23 15:35:42
Item downloaded from	<a href="https://hdl.handle.net/10468/14493">https://hdl.handle.net/10468/14493</a>

Ollscoil na hÉireann, Corcaigh  
**National University of Ireland, Cork**



Development of a ribosome profiling protocol to study translation  
in the yeast *Kluyveromyces marxianus*

Thesis presented by  
**Darren Fenton, BSc**  
for the degree of  
**Doctor of Philosophy**

**University College Cork**  
**School of Microbiology**

Head of School/Department: Prof Paul O'Toole  
Supervisors: Prof. Pavel Baranov & Prof. John  
Morrissey

2022

## Table of Contents

<b><i>Declaration.....</i></b>	<b><i>3</i></b>
<b><i>Abstract.....</i></b>	<b><i>4</i></b>
<b><i>Abbreviations .....</i></b>	<b><i>6</i></b>
<b><i>Acknowledgements .....</i></b>	<b><i>7</i></b>
<b><i>Manuscripts Submitted or Published during this PhD.....</i></b>	<b><i>8</i></b>
<b><i>Chapter 1 - Part I: Translation.....</i></b>	<b><i>9</i></b>
<b><i>Chapter 1 Part II: Kluyveromyces marxianus .....</i></b>	<b><i>23</i></b>
<b><i>Aims and Objectives of this thesis.....</i></b>	<b><i>35</i></b>
<b><i>Chapter 2 - Development of a Ribosome Profiling Protocol to Study Translation in Kluyveromyces marxianus .....</i></b>	<b><i>36</i></b>
<b><i>Chapter 3 - Integrated data-driven reannotation of the Kluyveromyces marxianus genome reveals an expanded protein coding repertoire.....</i></b>	<b><i>58</i></b>
<b><i>Chapter 4 - Heat-shock Induces a Rapid Increase in the Genes Involved in Cellular Respiration in the Yeast Kluyveromyces marxianus.....</i></b>	<b><i>112</i></b>
<b><i>Chapter 5 – General Discussion and Future Research .....</i></b>	<b><i>133</i></b>

## **Declaration**

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism.

Signed..... Darren Fenton



## Abstract

During mRNA translation, the ribosome protects ~28 nt of mRNA within its mRNA tunnel. Ribosome profiling is a method which takes advantage of this protected mRNA fragment, commonly referred to as the ribosome protected fragment (RPF). This method uses endonucleases to digest unprotected mRNA, purifying the RPF and generating a cDNA library. Deep sequencing of these cDNA libraries can reveal the locations of all translating ribosomes *in vivo*. These data can be used to study aspects of mRNA translation including recoding (e.g. +1 frameshifting), ribosome stalling and differential gene expression between multiple conditions. In addition, as RPFs can be mapped back to the genome, novel translated ORFs may be discovered including novel protein coding genes and regulatory elements of mRNA translation such as upstream open reading frames present in the 5' leaders (uORFs). Ribosome profiling was initially developed in the model yeast *S. cerevisiae*, but has since spread to human and bacterial models.

The first results chapter of this thesis describes the development of a ribosome profiling protocol to study translation in the yeast *Kluyveromyces marxianus*, which has not previously been developed. This protocol includes detailed steps to carry out the wet lab protocol as well as some aspects on computational processing. This protocol is accompanied by the release of the *K. marxianus* genome and transcriptome to a publicly available genome (GWIPS-viz) and transcriptome browser (Trips-Viz) which are tailored specifically for ribosome profiling data. Together, these protocols and browsers will allow others in the *K. marxianus* community to generate ribosome profiling data and upload/analyse data via these genome browsers.

The second results chapter describes the use of a combination of multiple methods including ribosome profiling, RNA-seq and transcript start site sequencing in what is described as “multiomics”. Using these multiomic data, the transcriptional and translational landscape is explored revealing a wide range of features including N-terminal extensions, upstream open reading frames and frameshifting. In addition, these data were used to generate a more complete and accurate genome annotation for *K. marxianus* by incorporating previously unannotated genes as well as a large number of gene annotation corrections such as splicing errors and start codon corrections.

The third results chapter describes using the developed ribosome profiling protocol to study how *K. marxianus* adapts to a rapid increase in temperature in an effort to understand how thermotolerant yeast adapts to such a stress. These data include both ribosome profiling and RNA-seq with multiple timepoints. Interestingly, the response to heat shock included a large and rapid response whereby cellular respiration is immediately upregulated, supported by both ribosome profiling, RNA-sequencing and biochemical assays.

## Abbreviations

Abbreviation	Term
aa	Amino Acid
ATP	Adenosine Triphosphate
aORF	Antisense Open Reading Frame
CDS	Coding DNA Sequence
GDP	Guanosine Diphosphate
GTP	Guanosine Triphosphate
GO	Gene Ontology
GWIPS-viz	Genome-wide Information on Protein Synthesis Visualized
iORF	Internal Open Reading Frame
MTS	Mitochondrial Targeting Signal
NCC	Near Cognate Codon
NCY	Non-Conventional Yeasts
NGS	Next Generation Sequencing
NTE	N-terminal Extension
nt	Nucleotide
ORF	Open Reading Frame
PANT	Proteins with Alternative N-terminus
PAS	Polyadenylation Site
Ribo-Seq	Ribosome Profiling
RNA-Seq	RNA Sequencing
RPF	Ribosome Protected Fragment
TE	Translation Efficiency
TC	Ternary Complex
Trips-Viz	Transcriptome-wide Information on Protein Synthesis Visualized
TSS	Transcript Start Site
uORF	Upstream Open Reading Frame
WGD	Whole Genome Duplication

## Acknowledgements

The work presented in this thesis was funded as part of the CHASSY project. This project received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation Grant Agreement No. 720824.

I sincerely thank my supervisors Prof. John Morrissey and Prof. Pasha Baranov for allowing me to pursue a PhD and carry out research in their labs, and the opportunities this PhD has provided me. Many long insightful discussions sharing their immense knowledge and allowing me to pursue my own and shared ideas kept me encouraged all the way to the end.

I thank Martina Yordanova for patiently showing me the ropes during my first days and weeks of lab work and her constant technical advice over the course of my PhD. I also thank Gary Loughran and Sinead O'Loughlin of the Recode lab who also shared their technical expertise along the way.

My thanks go to my fellow members of the LAPTI and Recode labs, past and present, Stephen H, Paddy, Janinah, Jack, Alla, Kate, Aoife, Oscar and James who always made the office and lab a fun and sociable place to work and in particular, Patrick, Audrey and Stephen K for guiding me through bioinformatics. I thank Pramodo, Luke and Oza for constant encouragement and humour. I thank past and present members of John's lab especially Javier, Angela, Noemi, Arun for all their valuable advice relating to everything yeast and the occasional pint at Tom Barry's. I thank Michał, Masha and Joanna (University of Warsaw) for giving me the opportunity to carry out research in Warsaw in what was truly a different experience and one I will not forget. I thank Pat Allen, Noreen Casey and other technical staff for keeping the Biochemistry floor of the WGB running, especially for keeping our aging ultracentrifuges running which were crucial for my work and others. I would like to thank all my collaborators in the CHASSY project for their insights and helpful discussions at the annual meetings.

Finally, I thank Roisín, my family, friends, Casper and my feline friends Klocek, Kicha, Miles and Peachu.

## Manuscripts Submitted or Published during this PhD

Montini N, Doughty TW, Domenzain I *et al.* Identification of a novel gene required for competitive growth at high temperature in the thermotolerant yeast *Kluyveromyces marxianus*. *Microbiology* 2022;**168**, DOI: 10.1099/mic.0.001148.

Darren A Fenton, Stephen J Kiniry, Martina M Yordanova, Pavel V Baranov, John P Morrissey, Development of a Ribosome Profiling Protocol to Study Translation in *Kluyveromyces marxianus* , *FEMS Yeast Research*, 2022;, foac024, <https://doi.org/10.1093/femsyr/foac024>

Fenton DA, Świrski M, O'Connor PBF *et al.* Integrated data-driven reannotation of the *Kluyveromyces marxianus*; genome reveals an expanded protein coding repertoire. *bioRxiv* 2022:2022.03.25.485750.

Coral-Medina A, Fenton DA, Varela J *et al.* Regulated use of alternative Transcription Start Sites controls the production of cytosolic or mitochondrial forms of branched-chain aminotransferase in *Kluyveromyces marxianus*; *bioRxiv* 2022:2022.04.27.489738.

Coral-Medina A, Fenton DA, Varela J *et al.* ASP3 evolution and role in asparagine consumption in *Saccharomyces* sp. *FEMS Yeast Research*, 2022; <https://doi.org/10.1093/femsyr/foac044>

# Chapter 1 - Part I: Translation

## Translation

Translation is the process of decoding a messenger RNA (mRNA) into a protein, which is carried out by the ribosome. The ribosome is a large ribozyme consisting of two subunits, the 40S and 60S. When the two subunits are bound upon translation initiation, they form an 80S ribosome, responsible for synthesising a polypeptide chain. A ribosome is described as having 3 sites (exit-site, peptidyl-site and aminoacyl-site (or E, P and A)) describing transfer RNA (tRNA) positioning inside the ribosome. The A-site is designated for incoming tRNAs, the P-site contains the already decoded tRNA while the E-site allows exit of tRNAs from the ribosome.

Translation is the biggest biological process occurring in the cell, by both energy consumption and abundance of factors involved. A yeast cell is estimated to contain 200,000 ribosomes and between 15,000-16,000 mRNAs and of these, ~33% encode ribosome proteins (Warner 1999; Zenklusen, Larson and Singer 2008). Yeast cells growing in log-phase in rich medium are estimated to produce almost 13,000 protein per second (von der Haar 2008), which is considered to be limited by the number of ribosomes available (Shah et al. 2013). In addition, translation elongation factors are among the most abundant proteins in a yeast cell. Translation is broken into three parts described below.

## Initiation

During the first step of translation, known as translation initiation, a scanning ribosome recognizes a start codon and is primed for elongation. Before a ribosome can recognize a start codon, a number of steps is first required. The first step of initiation involves the assembly of a ternary complex consisting of the eukaryotic initiation factor 2 protein (eIF2) bound to Guanosine triphosphate (GTP) and Met-tRNA<sub>i</sub><sup>Met</sup>. The ternary complex (TC), 40S ribosome small subunit and other initiation factors (eIF1A, eIF3 and eIF5) come together to form a 43S preinitiation complex (PIC). The eIF4F complex of initiation factors is responsible for preparing the mRNA for translation before a 43S ribosome can arrive. In this complex, eIF4E will bind the 5' mRNA cap forming a scaffold of eIF4A, eIF4B and eIF4G. This eIF4F complex will bind to the 43S PIC to form a 48S PIC. It is this 48S PIC that will move along the 5' leader of mRNA and recognize a start codon. Once the PIC recognizes a start codon,

the initiation factors become released and the large 60S subunit binds to the 40S subunit to form the 80S ribosome.

The canonical start codon for translation initiation is AUG, however, this is not always the case. For example, the *S. cerevisiae* gene *GRS1* provided one of the first examples of non-AUG initiation within a functional protein-coding gene. *GRS1* encodes a glycyl tRNA synthetase where ribosomes initiate at an AUG or UUG start codon. The AUG start codon encodes the cytoplasmic isoform while upstream translation initiation at UUG creates an N-terminal extension that includes a mitochondrial targeting signal (Chang and Wang 2004). These non-AUG start codons are commonly referred to as NCC start codons (near-cognate codons).

A major translation control pathway in yeast involves the TC component eIF2a and is commonly referred to as the integrated stress response (ISR). The yeast kinase Gcn2 (general control nonrepressed 2) can phosphorylate eIF2a which prevents its GTP binding capability, in turn this shuts down translation globally while increasing translation of the gene *GCN4* (Dever *et al.* 1992; Hinnebusch 2005). Gcn2 is a major transcription factor responsible for activation of amino acid biosynthesis genes. Thus, amino acid starvation is a major activator of eIF2a phosphorylation, with Gcn2 binding to increasing uncharged tRNA concentrations. Other stresses which activate the ISR include glucose starvation (Yang, Wek and Wek 2000) and high concentrations of sodium chloride (Goossens *et al.* 2001) and boron (Ulusik *et al.* 2011). While the ISR regulates global translation initiation in the cell, there are well studied examples of translation initiation regulation of specific genes including *GCN4*, *CPA* and *HAC1*.

One of the most well characterized genes to be translationally regulated in yeast is General control transcription factor (*GCN4*), which has a homologue, Activating Transcription Factor 4 (*ATF4*), in higher eukaryotes. This gene, regulated via translation initiation, encodes a transcription factor responsible for transcriptional activation of genes involved in amino acid biosynthesis pathways. Under “normal” or “rich” growth conditions when amino acids are abundant, the translation of *GCN4* main coding DNA sequence (CDS) is repressed by 4 upstream open reading frames (uORFs) embedded in the long ~591 mRNA 5' leader (Hinnebusch 1984). In this model, translation initiation occurs at uORF1, producing a tripeptide. However, it is estimated that ~50% of ribosomes remain bound to the mRNA after

termination. These mRNA-bound ribosomes are suggested to have a 40S subunit with the retention of a few initiation factors including eIF3 (Szamecz *et al.* 2008). However, as TCs are abundant, these 40S subunits become primed for translation and reinitiate at uORFs 2, 3 and 4. During amino acid starvation, the Gcn2 kinase phosphorylates a serine residue (Ser51) on eIF2a. This phosphorylation of eIF2a inhibits Guanosine Diphosphate (GDP) to Guanosine Triphosphate (GTP) recycling thus reducing TC levels, which triggers a massive translation initiation reduction in the cell. Due to reduced TC levels, ribosomes migrating across the mRNA and uORFs are less likely to receive available TC. It is important to note that a 40S bound to a TC is not necessary for ribosome-mRNA interactions which allow migration. However, between the last uORF and the main CDS, ribosomes are more likely to become bound to TC and translation of the *GCN4* main CDS is enhanced (Hinnebusch 2005). This model also highlights the more critical roles of uORFs 1 and 4 as uORF1 translation favours ribosomes remaining bound to mRNA post termination while uORF 4 favours release of most ribosomes after termination. It has also been suggested that nucleotide sequences both 5' and 3' of uORF1 promotes this retention of ribosomes post termination (Grant and Hinnebusch 1994).

*CPA1* (Carboxypeptidase A1) encodes a subunit of carbamoyl phosphate synthetase involved in the arginine biosynthesis pathway. Early studies elucidated that *CPA1* was negatively regulated by the presence of arginine in growth medium (Thuriaux *et al.* 1972). Later work would find that a uORF peptide encoded the 5' mRNA leader is essential for translational repression under arginine conditions (Werner *et al.* 1987). This uORF encodes a 25 amino acid peptide which stalls the ribosome when arginine levels are high and is conserved across the fungi kingdom (Hood, Spevak and Sachs 2007). This 25 aa peptide is generally known as the arginine attenuator peptide (AAP). In this regulation mechanism, the presence of high arginine concentrations stalls ribosomes at the uORF stop codon. This stalling blocks scanning ribosomes and thus no initiation occurs at the *CPA1* main CDS. Therefore, when arginine levels drop, *CPA1* translation is increased, leading to an increase in arginine biosynthesis. In addition to ribosome stalling, NMD (Nonsense mediated decay), which detects and destroys aberrant mRNAs with premature stop codons (Kervestin and Jacobson 2012), also plays a role. As ribosome stalling occurs at a uORF stop codon, the NMD quality control pathway recognizes CPA1 mRNA and triggers destruction of *CPA1* mRNA.



In yeast, the unfolded protein response (UDR) allows cells to adapt to range of stresses including heat shock. The Hac1 (Homologous to Atf/Creb1) protein functions as a transcription factor responsible for triggering the UDR upon stress (Nikawa, Hosaka and Yamashita 1993; Cox and Walter 1996; Mori *et al.* 1996). In unstressed cells, this mRNA contains an intron, a sequence of mRNA usually removed by splicing catalysed by the spliceosome complex. When this intron is present as part of the mRNA, it hybridizes to a region in the 5' leader and represses translation initiation, reducing the Hac1 protein levels in unstressed cells. However, during stress, the intron is removed and the translation efficiency of the HAC1 mRNA increases (Cox and Walter 1996; Chapman and Walter 1997), which will result in the activation of the UDR transcriptional programme.

### **Elongation**

Once the ribosome has recognized the start codon and initiation is complete, the ribosome is now ready for the process of multiple rounds of elongation, one codon at a time. Elongation involves the eukaryotic elongation factors (eEF) eEF1A (encoded by *TEF1* and *TEF2*), eEF1B and eEF2. With the P-site occupied by Met-tRNA<sup>Met</sup> and the A-site unoccupied, a complex of eEF1A-GTP-aa-tRNA binds to the A-site of the ribosome, if the tRNA is paired with the codon, GTP hydrolysis triggers the release of eEF1A-GDP out of the ribosome leaving tRNA-aa inside the A-site. A rapid peptide bond formation in the peptidyl transferase centre (PTC) adds the second amino acid to the first methionine amino acid. Once this peptide bond is formed, the ribosome rachets moving the two tRNAs into P/E and A/P states with the tRNA acceptor stems in the E and P site. The second elongation factor eEF2-GTP promotes the translocation of the tRNAs individually into the E and P sites. The deacylated tRNA is released from the E-site. The now new P-site tRNA remains bound to the two amino acid peptide chain and elongation is ready for its second round, which will begin when the next correct eEF1A-GTP-aa-tRNA enters the A-site. Recycling of eEF1A-GDP to eEF1A-GTP is carried out by eEF1B.

While the eEF1A, eEF1B and eEF2 translation factors are conserved among eukaryotes, fungi contain an essential third elongation factor named eEF3. However one bioinformatic study suggested that eEF3 is present more widely in the genomes of unicellular organisms including *S. cerevisiae* (Mateyak *et al.* 2018), whose function in translation has only recently been answered. Ranjan *et al.*, 2021 suggested eEF3 plays a critical role during the

translocation step of the elongation cycle. Here, eEF3 functions to accelerate the E-site tRNA release from the ribosome during mRNA-tRNA translocation by inducing the L1 stalk to adopt a conformation favouring tRNA release. Strengthening their findings with ribosome profiling, a yeast strain carrying an eEF3 depletion resulted in the overrepresentation of 28 nucleotide (nt) footprints, similar to the effects of cycloheximide treatment which is widely considered to block E-site tRNA ejection. Originally identified as the initiation factor eIF5A, due to its stimulation of the first peptide bond between Met-tRNA and puromycin (Kemper, Berry and Merrick 1976; Schreier, Erni and Staehelin 1977; Benne and Hershey 1978), eIF5A was later shown to promote translation elongation by promoting the Proline-Proline (Pro-Pro) bond formation. More recently eIF5A has been implicated to play a broader role in translation, by promoting the elongation of a wide range of stalling sequences identified by ribosome profiling including both poly-Pro and non-poly-Pro sequences (Schuller *et al.* 2017) and stimulating eRF1-mediated peptidyl-tRNA hydrolysis during translation termination (Schuller *et al.* 2017). eIF5A is also known for its rare hypusine modification which is critical for its activity (Park *et al.* 1991, 2011; Saini *et al.* 2009).

## Termination

In the final step of ribosome translation, a ribosome reaches a stop codon at UAA, UGA or UAG, and the ribosome will terminate and disassemble through ribosome release factors. This is important to allow not only allow the release of the synthesized peptide, but for recycling of the ribosome 40S and 60S subunits to continue additional rounds of translation. For this to occur, a ribosome elongates to the end of a CDS and a stop codon is recognized in the A-site by eRF1 (encoded by *SUP45*). eRF1 is composed of three separate functional parts. The N-terminus of eRF1 is responsible for recognition of the stop codon (Bertram *et al.* 2000). The central domain is responsible for efficient hydrolysis of the nascent peptide bound to the P-site tRNA, centred around a critical methylated glycine-glycine-glutamine (GGQ) motif (Heurgué-Hamard *et al.* 2005) and the C-terminus interacts to eRF3, which is critical for correct stop codon recognition (Wada and Ito 2014). eRF1 interacts with GTP-bound eRF3 (encoded by *SUP35*), of which the GTP hydrolysis of eRF3-GTP is critical for correct stop codon recognition (Salas-Marco and Bedwell 2004) and catalysing peptide release (Eyler, Wehner and Green 2013). Upon this GTP hydrolysis, eRF3 dissociates, leaving eRF1 in the A-site. Next, the ATPase Rli1 (also known as ABCE1) enters the A-site and interacts with eRF1 catalysing the efficient hydrolysis of the aminoacyl bond between the tRNA and the

synthesized peptide. It is the interaction of eRF1 and Rli1 which stimulates the GGQ motif of eRF1 to move towards the peptidyl-transferase centre of the ribosome and promote hydrolysis of the nascent peptide, however this activity is not dependent on ATPase activity of Rli1 (Khoshnevis *et al.* 2010; Shoemaker and Green 2011). The 80S ribosome now has a deacylated tRNA in the P-site. It is now the ATPase activity of Rli1 catalyses a ribosome conformation change which allows the dissociation of ribosome subunits, which then become bound by initiation factors to continue another round of mRNA translation (Pisarev, Hellen and Pestova 2007). Depletion of Rli1 in yeast allows ribosome reinitiation within the 3' trailer of the mRNA (Young *et al.* 2015). Thus Rli1 also functions in both ribosome termination and ribosome recycling. More recently, the initiation factor eIF5A has been proposed to play a critical role in translation termination by stimulating the rate of eRF1-mediated peptidyl-tRNA hydrolysis by ~17-fold (Schuller *et al.* 2017).

### **Frameshifting**

While mRNA decoding was originally thought to occur in a single open reading frame, there have been many examples where during decoding of an mRNA, a ribosome can shift (frameshift) to the +1 or -1 frame, producing an alternative protein. These events belong to a process known as recoding (Gesteland, Weiss and Atkins 1992). This section will focus on +1 frameshifting in yeast.

To allow efficient +1 frameshifting in yeast, a minimum seven nucleotide heptamer encoding a particular set of codons must be present, two in the 0-frame and one codon in the +1-frame (XXX\_YYY\_Z). The P-site codon (XXX) must contain a tRNA considered to be “slippery”, allowing repositioning of the ribosome to a +1 frame. The A-site codon (YYY) must be slow-to-decode due to low intracellular levels of charged tRNA for that codon, should this codon have a high abundance tRNA, frameshifting efficiency would be expected to reduce. The +1 codon (YYZ) must have a higher abundance of charged tRNA relative to the P-site to allow efficient recognition and continued elongation in the +1 frame. Therefore a combination of weak P-site codon-anticodon pairing and a high ratio of +1 codon decoding tRNAs to 0-frame codon directly stimulate frameshifting.

The retroviral transposable element *Ty1* was the first example in yeast employing +1 frameshifting (Clare, Belcourt and Farabaugh 1988; Belcourt and Farabaugh 1990). In the *Ty1* frameshift heptamer, a ribosome with weak CUU codon-anticodon tRNA-mRNA pairing

in the P-site repositions to the +1 frame followed by an abundant incoming tRNA in the +1 frame. The 0-frame CDS encodes Gag, the structural protein of the viral particle (capsid) and the +1 frame CDS encodes Pol which has reverse transcriptase activity. It has been shown that changing the ratio of Gag to Gag-Pol expression can severely reduce Ty1 transposition frequency (Xu and Boeke 1990; Kawakami *et al.* 1993), suggesting +1 frameshifting plays an important role in the functional ratio of 0-frame to full length protein production. This Ty1 frameshift heptamer was later discovered within the coding region of *ABP140*, an actin binding protein, responsible for 3-methylcytidine (m3C) modification at position 32 of threonine and serine tRNAs (Asakura *et al.* 1998; Noma *et al.* 2011).

Studying the effects of all 64 P-site codons in the CUU\_AGG\_C heptamer, the most frameshift prone P-site codons were as follows, CUU>CCG>GCG>GGG, therefore the Ty1 heptamer employs the most efficient P-site (CUU) codon to allow efficiency frameshifting (Vimaladithan and Farabaugh 1994). The discovery of other yeast genes employing +1 frameshifting were later discovered in Ty3 (Farabaugh, Zhao and Vimaladithan 1993), *EST3* (subunit of telomerase) (Morris and Lundblad 1997) and *OAZ1* (Ornithine decarboxylase antizyme 1) (Palanimurugan *et al.* 2004, reviewed in the following section). The most recent identification of a novel +1 frameshifting was identified in the upstream open reading frame (uORF) of *YSF1* (Ivanov *et al.* 2020).

Polyamines are positively charged low weight diamines and polyamines including putrescine, spermidine and spermine which are present in all cells (Wallace 2009). Translation of the *OAZ1* mRNA is responsible for regulating intracellular polyamine concentrations using polyamine stimulated +1 frameshifting and this mechanism is conserved from yeast to humans (Matsufuji *et al.* 1996; Palanimurugan *et al.* 2004; Ivanov and Atkins 2007). The *OAZ1* mRNA contains two translated CDS regions, the 0-frame ORF encodes ornithine decarboxylase (*odc1*) which is responsible for the conversion of ornithine to putrescine. However, translation of this 0-frame product increases intracellular polyamine levels to a point whereby +1 frameshifting at the 0-frame stop codon is induced. The ribosome (now in the +1 frame) translates the antizyme CDS which binds and inhibits ornithine decarboxylase, therefore providing an autoregulation mechanism. In addition to inhibiting the enzymatic activity of *odc1*, antizyme also accelerates the degradation of *odc1* (Beenukumar *et al.* 2015).

## Global studies of Translation in Yeast

To study gene expression genome-wide in yeast, microarrays and in the last decade, RNA sequencing (RNA-seq), have become common methods to study environmental changes. However, these methods have solely focused on mRNA abundances between samples, revealing only transcriptional changes, ignoring translation control of gene expression (Sonenberg and Hinnebusch 2009). Some milestones have been achieved with genome-wide studies of translation, most notably from a combination of microarray analyses with polysome profiling, whereby mRNAs can be separated depending on the number of ribosomes per mRNA (called polysomes) (Arava *et al.* 2003). These experiments provided data such as ribosome densities ranging from 0.03-3.3 ribosomes per 100 nucleotides and provided evidence of translational control to known examples such as *GCN4* and *HAC1* (Arava *et al.* 2003). A major breakthrough for genome-wide studies in translation came in 2009 with the advent of ribosome profiling (Ingolia *et al.* 2009).

## Ribosome Profiling

Ribosome profiling, first described in yeast in 2009, provides a genome-wide (or transcriptome-wide, depending on where data is aligned to the transcriptome or genome) view of translation with codon-resolution (Ingolia *et al.* 2009). During mRNA translation, the ribosome protects a fragment of mRNA within the mRNA tunnel from endonuclease digestion (Wolin and Walter 1988). Ribosome profiling is a technique which purifies these ribosome protected fragments (RPFs). Purified RPFs are then converted to a cDNA library suitable for deep sequencing (usually with Illumina sequencing platforms, e.g. HiSeq 4000), revealing the locations of ribosomes genome-wide.

Ribosome profiling typically involves the following steps. First, ribosomes are paused *in vivo* using ultra cold temperatures and a translation inhibitor, usually cycloheximide. Cells are then mechanically broken and the lysate is clarified, containing total RNA, polysomes and other smaller cellular components. This lysate is then treated with an endonuclease to digest all unprotected RNA, during which ribosome protected fragments (RPFs) are generated. Different exonucleases are employed in different organisms with RNase I being the most widely used in yeast and human studies. The lysate is loaded onto a sucrose gradient and ultra-centrifuged to separate out the lysate components by size, optimized for collection of 80S monosomes. With the 80S monosome isolated, the ribosome is denatured and the ~30nt

RPF is released. This RPF is then size-selected on a polyacrylamide gel with the aid of RNA markers, purified and a cDNA library is prepared. After sequencing, RPFs can be mapped to the genome or transcriptome interest to provide a global view of translation.

As RNase I typically digests mRNA up to 12 nt from the 5' end of mRNA to the first nucleotide of the P-site codon, ribosome profiling data displays triplet (or codon) periodicity, which reflects to codon-wise movement of ribosomes along mRNAs. Therefore it is possible to determine which reading frame is being translated due to the periodic signal. In yeast, ribosome profiling alone has increased the our known size and diversity of the translome. In *S. cerevisiae*, ribosome profiling has expanded the number of known translated upstream open reading frames (Ingolia *et al.* 2009; Brar *et al.* 2012; Spealman *et al.* 2018), and has also revealed much greater diversity of the N-terminal translation due to non-canonical translation upstream of mRNAs, of which many encode mitochondrial targeting signals (Monteuuis *et al.* 2019). Ribosome profiling has also been employed extensively to study differential gene expression in various conditions. RNA-seq is commonly carried out in parallel to ribosome profiling as a control to determine whether a gene is controlled transcriptionally or translationally. In *S. cerevisiae*, these experiments have revealed a landscape of both transcriptional and translational control in many pathways and responses such as meiosis (Brar *et al.* 2012), oxidative stress (Gerashchenko, Lobanov and Gladyshev 2012; Blevins *et al.* 2019) and heat shock (Mühlhofer *et al.* 2019). In addition, ribosome profiling has also been carried out in other yeast species including *Saccharomyces paradoxus* (McManus *et al.* 2014), *Schizosaccharomyces pombe* (Duncan and Mata 2014), *Saccharomyces uvarum* (Spealman *et al.* 2018), *Komagataella phaffii* (Alva, Riera and Chartron 2021), *Candida albicans* (Sharma *et al.* 2021), and, as reported later in this thesis, *Kluyveromyces marxianus* (Fenton *et al.* 2022).

## References

- Alva TR, Riera M, Chartron JW. Translational landscape and protein biogenesis demands of the early secretory pathway in *Komagataella phaffii*. *Microb Cell Fact* 2021;**20**:19.
- Arava Y, Wang Y, Storey JD *et al*. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*; *Proc Natl Acad Sci* 2003;**100**:3889 LP – 3894.
- Asakura T, Sasaki T, Nagano F *et al*. Isolation and characterization of a novel actin filament-binding protein from *Saccharomyces cerevisiae*. *Oncogene* 1998;**16**:121–30.
- Beenukumar RR, Gödderz D, Palanimurugan R *et al*. Polyamines directly promote antizyme-mediated degradation of ornithine decarboxylase by the proteasome. *Microb cell (Graz, Austria)* 2015;**2**:197–207.
- Belcourt MF, Farabaugh PJ. Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. *Cell* 1990;**62**:339–52.
- Benne R, Hershey JW. The mechanism of action of protein synthesis initiation factors from rabbit reticulocytes. *J Biol Chem* 1978;**253**:3078–87.
- Bertram G, Bell HA, Ritchie DW *et al*. Terminating eukaryote translation: domain 1 of release factor eRF1 functions in stop codon recognition. *RNA* 2000;**6**:1236–47.
- Blevins WR, Tavella T, Moro SG *et al*. Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. *Sci Rep* 2019;**9**:11005.
- Brar GA, Yassour M, Friedman N *et al*. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* 2012;**335**:552–7.
- Chang K-J, Wang C-C. Translation initiation from a naturally occurring non-AUG codon in *Saccharomyces cerevisiae*. *J Biol Chem* 2004;**279**:13778–85.
- Chapman RE, Walter P. Translational attenuation mediated by an mRNA intron. *Curr Biol* 1997;**7**:850–9.
- Clare JJ, Belcourt M, Farabaugh PJ. Efficient translational frameshifting occurs within a conserved sequence of the overlap between the two genes of a yeast Ty1 transposon. *Proc Natl Acad Sci U S A* 1988;**85**:6816–20.
- Cox JS, Walter P. A novel mechanism for regulating activity of a transcription factor that controls the unfolded protein response. *Cell* 1996;**87**:391–404.
- DE E, KA W, Green R. Eukaryotic release factor 3 is required for multiple turnovers of peptide release catalysis by eukaryotic release factor 1. *J Biol Chem* 2013;**288**:29530–8.
- Dever TE, Feng L, Wek RC *et al*. Phosphorylation of initiation factor 2 alpha by protein

- kinase GCN2 mediates gene-specific translational control of GCN4 in yeast. *Cell* 1992;**68**:585–96.
- Duncan CDS, Mata J. The translational landscape of fission-yeast meiosis and sporulation. *Nat Struct Mol Biol* 2014;**21**:641–7.
- Farabaugh PJ, Zhao H, Vimaladithan A. A novel programmed frameshift expresses the POL3 gene of retrotransposon Ty3 of yeast: frameshifting without tRNA slippage. *Cell* 1993;**74**:93–103.
- Fenton DA, Kiniry SJ, Yordanova MM *et al.* Development of a Ribosome Profiling Protocol to Study Translation in the yeast *Kluyveromyces marxianus*. *bioRxiv* 2022:2022.02.06.478964.
- Gerashchenko M V, Lobanov A V, Gladyshev VN. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc Natl Acad Sci* 2012;**109**:17394 LP – 17399.
- Gesteland RF, Weiss RB, Atkins JF. Recoding: reprogrammed genetic decoding. *Science* 1992;**257**:1640–1.
- Goossens A, Dever TE, Pascual-Ahuir A *et al.* The protein kinase Gcn2p mediates sodium toxicity in yeast. *J Biol Chem* 2001;**276**:30753–60.
- Grant CM, Hinnebusch AG. Effect of sequence context at stop codons on efficiency of reinitiation in GCN4 translational control. *Mol Cell Biol* 1994;**14**:606–18.
- von der Haar T. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. *BMC Syst Biol* 2008;**2**:87.
- Heurgué-Hamard V, Champ S, Mora L *et al.* The Glutamine Residue of the Conserved GGQ Motif in *Saccharomyces cerevisiae* Release Factor eRF1 Is Methylated by the Product of the YDR140w Gene\*. *J Biol Chem* 2005;**280**:2439–45.
- Hinnebusch AG. Evidence for translational regulation of the activator of general amino acid control in yeast. *Proc Natl Acad Sci U S A* 1984;**81**:6442–6.
- Hinnebusch AG. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol* 2005;**59**:407–50.
- Hood H, Spevak C, Sachs M. Evolutionary changes in the fungal carbamoyl-phosphate synthetase small subunit gene and its associated upstream open reading frame. *Fungal Genet Biol* 2007;**44**:93–104.
- Ingolia NT, Ghaemmaghami S, Newman JRS *et al.* Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 2009;**324**:218–



23.

- Ivanov IP, Atkins JF. Ribosomal frameshifting in decoding antizyme mRNAs from yeast and protists to humans: close to 300 cases reveal remarkable diversity despite underlying conservation. *Nucleic Acids Res* 2007;**35**:1842–58.
- Ivanov IP, Gaikwad S, Hinnebusch AG *et al.* Conserved +1 translational frameshifting in the <em>S. cerevisiae</em> gene encoding YPL034W. *bioRxiv* 2020:2020.04.29.069534.
- Kawakami K, Pande S, Faiola B *et al.* A rare tRNA-Arg(CCU) that regulates Ty1 element ribosomal frameshifting is essential for Ty1 retrotransposition in *Saccharomyces cerevisiae*. *Genetics* 1993;**135**:309–20.
- Kemper WM, Berry KW, Merrick WC. Purification and properties of rabbit reticulocyte protein synthesis initiation factors M2B $\alpha$  and M2B $\beta$ . *J Biol Chem* 1976;**251**:5551–7.
- Kervestin S, Jacobson A. NMD: a multifaceted response to premature translational termination. *Nat Rev Mol Cell Biol* 2012;**13**:700–12.
- Khoshnevis S, Gross T, Rotte C *et al.* The iron–sulphur protein RNase L inhibitor functions in translation termination. *EMBO Rep* 2010;**11**:214–9.
- Mateyak MK, Pupek JK, Garino AE *et al.* Demonstration of translation elongation factor 3 activity from a non-fungal species, *Phytophthora infestans*. *PLoS One* 2018;**13**:e0190524.
- Matsufuji S, Inazawa J, Hayashi T *et al.* Assignment of the human antizyme gene (OAZ) to chromosome 19p13.3 by fluorescence in situ hybridization. *Genomics* 1996;**38**:102–4.
- McManus CJ, May GE, Spealman P *et al.* Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res* 2014;**24**:422–30.
- Monteuuis G, Miścicka A, Świrski M *et al.* Non-canonical translation initiation in yeast generates a cryptic pool of mitochondrial proteins. *Nucleic Acids Res* 2019;**47**:5777–91.
- Mori K, Kawahara T, Yoshida H *et al.* Signalling from endoplasmic reticulum to nucleus: transcription factor with a basic-leucine zipper motif is required for the unfolded protein-response pathway. *Genes Cells* 1996;**1**:803–17.
- Morris DK, Lundblad V. Programmed translational frameshifting in a gene required for yeast telomere replication. *Curr Biol* 1997;**7**:969–76.
- Mühlhofer M, Berchtold E, Stratil CG *et al.* The Heat Shock Response in Yeast Maintains Protein Homeostasis by Chaperoning and Replenishing Proteins. *Cell Rep*

- 2019;**29**:4593-4607.e8.
- Nikawa J, Hosaka K, Yamashita S. Differential regulation of two myo-inositol transporter genes of *Saccharomyces cerevisiae*. *Mol Microbiol* 1993;**10**:955–61.
- Noma A, Yi S, Katoh T *et al*. Actin-binding protein ABP140 is a methyltransferase for 3-methylcytidine at position 32 of tRNAs in *Saccharomyces cerevisiae*. *RNA* 2011;**17**:1111–9.
- Palanimurugan R, Scheel H, Hofmann K *et al*. Polyamines regulate their synthesis by inducing expression and blocking degradation of ODC antizyme. *EMBO J* 2004;**23**:4857–67.
- Park JH, Dias CAO, Lee SB *et al*. Production of active recombinant eIF5A: reconstitution in *E.coli* of eukaryotic hypusine modification of eIF5A by its coexpression with modifying enzymes. *Protein Eng Des Sel* 2011;**24**:301–9.
- Park MH, Wolff EC, Smit-McBride Z *et al*. Comparison of the activities of variant forms of eIF-4D. The requirement for hypusine or deoxyhypusine. *J Biol Chem* 1991;**266**:7988–94.
- Pisarev A V, Hellen CUT, Pestova T V. Recycling of eukaryotic posttermination ribosomal complexes. *Cell* 2007;**131**:286–99.
- Ranjan N, Pochopien AA, Chih-Chien Wu C *et al*. Yeast translation elongation factor eEF3 promotes late stages of tRNA translocation. *EMBO J* 2021;**40**:e106449.
- Saini P, Eyler DE, Green R *et al*. Hypusine-containing protein eIF5A promotes translation elongation. *Nature* 2009;**459**:118–21.
- Salas-Marco J, Bedwell DM. GTP hydrolysis by eRF3 facilitates stop codon decoding during eukaryotic translation termination. *Mol Cell Biol* 2004;**24**:7769–78.
- Schreier MH, Erni B, Staehelin T. Initiation of mammalian protein synthesis. I. Purification and characterization of seven initiation factors. *J Mol Biol* 1977;**116**:727–53.
- Schuller AP, Wu CC-C, Dever TE *et al*. eIF5A Functions Globally in Translation Elongation and Termination. *Mol Cell* 2017;**66**:194-205.e5.
- Sharma P, Wu J, Nilges BS *et al*. Humans and other commonly used model organisms are resistant to cycloheximide-mediated biases in ribosome profiling experiments. *Nat Commun* 2021;**12**:5094.
- Shoemaker CJ, Green R. Kinetic analysis reveals the ordered coupling of translation termination and ribosome recycling in yeast. *Proc Natl Acad Sci* 2011;**108**:E1392 LP-E1398.

- Sonenberg N, Hinnebusch AG. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* 2009;**136**:731–45.
- Spealman P, Naik AW, May GE *et al.* Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. *Genome Res* 2018;**28**:214–22.
- Szamecz B, Rutkai E, Cuchalová L *et al.* eIF3a cooperates with sequences 5' of uORF1 to promote resumption of scanning by post-termination ribosomes for reinitiation on GCN4 mRNA. *Genes Dev* 2008;**22**:2414–25.
- Thuriaux P, Ramos F, Piérard A *et al.* Regulation of the carbamoylphosphate synthetase belonging to the arginine biosynthetic pathway of *Saccharomyces cerevisiae*. *J Mol Biol* 1972;**67**:277–87.
- Ulusik I, Kaya A, Fomenko DE *et al.* Boron stress activates the general amino acid control mechanism and inhibits protein synthesis. *PLoS One* 2011;**6**:e27772.
- Vimaladithan A, Farabaugh PJ. Special peptidyl-tRNA molecules can promote translational frameshifting without slippage. *Mol Cell Biol* 1994;**14**:8107–16.
- Wada M, Ito K. A genetic approach for analyzing the co-operative function of the tRNA mimicry complex, eRF1/eRF3, in translation termination on the ribosome. *Nucleic Acids Res* 2014;**42**:7851–66.
- Wallace HM. The polyamines: past, present and future. *Essays Biochem* 2009;**46**:1–9.
- Warner JR. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci* 1999;**24**:437–40.
- Werner M, Feller A, Messenguy F *et al.* The leader peptide of yeast gene CPA1 is essential for the translational repression of its expression. *Cell* 1987;**49**:805–13.
- Wolin SL, Walter P. Ribosome pausing and stacking during translation of a eukaryotic mRNA. *EMBO J* 1988;**7**:3559–69.
- Xu H, Boeke JD. Host genes that influence transposition in yeast: the abundance of a rare tRNA regulates Ty1 transposition frequency. *Proc Natl Acad Sci U S A* 1990;**87**:8360–4.
- Yang R, Wek SA, Wek RC. Glucose limitation induces GCN4 translation by activation of Gcn2 protein kinase. *Mol Cell Biol* 2000;**20**:2706–17.
- Young DJ, Guydosh NR, Zhang F *et al.* Rli1/ABCE1 Recycles Terminating Ribosomes and Controls Translation Reinitiation in 3'UTRs In Vivo. *Cell* 2015;**162**:872–84.
- Zenklusen D, Larson DR, Singer RH. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol* 2008;**15**:1263–71.

## Chapter 1 Part II: *Kluyveromyces marxianus*

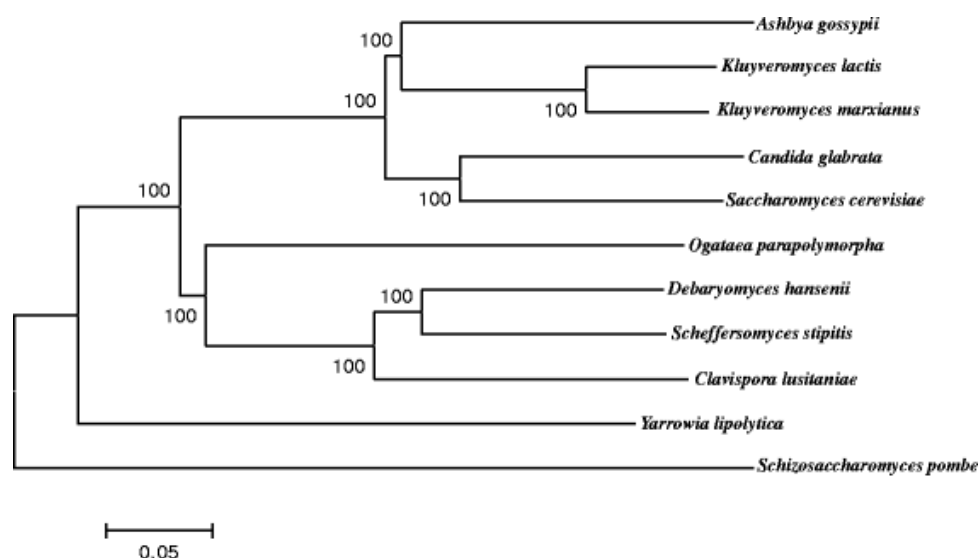
### *Yeast, Food and Biotechnology*

Yeasts are group of eukaryotic unicellular organisms which have an ancient history of food and beverage applications, the most notable example is the budding yeast *Saccharomyces cerevisiae*. *S. cerevisiae* has been widely used for bread baking, beer brewing and wine production since ancient times (Cavalieri *et al.* 2003; Sicard and Legras 2011; Shevchenko *et al.* 2014). In the modern age, *S. cerevisiae* is a common model organism to study eukaryotic processes via functional genomics (Botstein and Fink 2011). The full genome sequence of *S. cerevisiae* was completed in 1996 and a community organized effort produced a near-complete set of deletions for each gene (Goffeau *et al.* 1996; Winzeler *et al.* 1999). This yeast is also extensively employed in the biotechnology/food industry sector as a microbial cell factory to produce a range of valuable compounds including ethanol, oils and pharmaceuticals (Nandy and Srivastava 2018; Nielsen 2019; Parapouli *et al.* 2020).

While *S. cerevisiae* is undoubtedly the most widely studied and industrially relevant yeast, it is just one of over 1,000 known species in the budding yeast subphylum *Saccharomycotina* (Hittinger *et al.* 2015; Shen *et al.* 2018). Other yeasts have also garnered considerable interest in biotechnological applications due to potential advantages over *S. cerevisiae* for specific applications, and these species are collectively termed NCY yeasts (Non-Conventional Yeasts) (Rebello *et al.* 2018; Binati *et al.* 2021). The use of modern molecular approaches and modern genome sequencing for studying and engineering yeasts, these NCYs have become cell factories to produce a wide range of other valuable compounds such as ethanol and enzymes. Such examples of NCYs include *Yarrowia lipolytica*, an oleaginous yeast capable of utilizing alkanes and fatty acids, have made this species interesting as both a model for oleaginous yeast research and use in biotechnology (Desfougères *et al.* 2010; Gonçalves, Colen and Takahashi 2014). *Komagataella phaffii* (previously known as *Pichia pastoris*) has the ability to utilize methanol as a carbon source (methylophilic) and is often used as a protein expression system on both research and industrial scales (Zhu *et al.* 2019; Duman-Özdamar and Binay 2021; Yamada 2021). *Kluyveromyces lactis* is most widely known for its ability to consume lactose as a carbon source, a trait which *S. cerevisiae* does not possess (Lachance 1998; Schaffrath and Breunig 2000). Other notable examples of NCY include *Lachancea thermotolerans*, *Debaryomyces hansenii* and *Candida jadinii* (Binati *et al.* 2021).

### *Kluyveromyces marxianus* – Taxonomy, Physiology and Genome

In this introduction there will be a focus on *Kluyveromyces marxianus*, as this species is the subject of the work reported in this thesis. *K. marxianus* is another NCY which emerged as one such yeast with potential for industrial applications due a range of advantageous physiological traits discussed further below (Fonseca *et al.* 2008; Lane and Morrissey 2010; Morrissey *et al.* 2015; Varela *et al.* 2017). *K. marxianus* is a sister species of *K. lactis* and both species share the ability to consume lactose as a carbon source. These two species are the most commonly studied within the *Kluyveromyces* genus, although other species within the genus including *Kluyveromyces dobhanskii* and *Kluyveromyces aestuarii* have been identified and their genomes sequenced (Shen *et al.* 2018). The phylogenetic relationship between *K. marxianus* and other representative yeast species in the *Saccharomycotina* is presented in Figure 1.

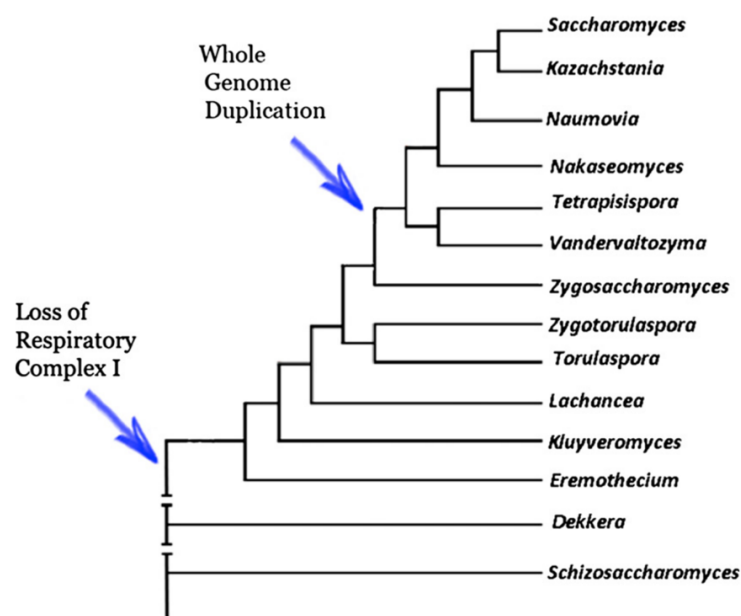


**Figure 1.** Phylogenetic tree of 11 yeasts based on 1361 concatenated amino acid sequences. This figure is taken from Lertwattanasakul *et al.*, 2015. Numbers at branch points represent bootstrap values, where 100 represents highest confidence.

*K. marxianus* is a thermotolerant yeast able to grow at high temperatures (up to 52°C) and is considered one of the fastest growing eukaryotic microbes by generation time (Groeneveld, Stouthamer and Westerhoff 2009). While *S. cerevisiae* is Crabtree positive (the cell favours fermentation in the presence of high glucose and oxygen concentrations), *K. marxianus* is Crabtree negative and does not undergo aerobic fermentation. A major physiological trait that distinguishes *K. marxianus* from *S. cerevisiae* is the ability to assimilate lactose as a carbon source (Lane *et al.* 2011; Carl *et al.* 2021). The presence of *LAC12* (encoding lactose

permease) and *LAC4* (encoding beta-galactosidase) in the genome allow *K. marxianus* to import lactose into the cell and metabolize the disaccharide to glucose and galactose, respectively. However, it is important to note that lactose uptake ranges significantly between strains due to variants of the *LAC12* gene (Varela *et al.* 2017). Other carbon sources include xylose and arabinose, and transporters which allow uptake of these pentose sugars have been recently identified in *K. marxianus* (Donzella *et al.* 2021).

*K. marxianus* has eight nuclear chromosomes excluding the ~45 kb mitochondrial chromosome, and the ~10.8 Mb genome encodes ~5100 proteins. Unlike *S. cerevisiae*, *K. marxianus* is a pre-Whole Genome Duplication (WGD) (Wolfe and Shields 1997) and therefore lacks the paralogous genes derived from the duplication event (see Figure 2). With modern high-throughput sequencing, a growing number of genome assemblies for *K. marxianus* have become available in the last decade, these include the genome assemblies of the strains KCTC 17555 (syn. CBS6556 and ATCC 26548), DMB1, NBRC 1777, IPE453 and DMKU3-1042 (Jeong *et al.* 2012; Suzuki, Hoshino and Matsushika 2014; Inokuma *et al.* 2015; Lertwattanasakul *et al.* 2015; Dasgupta *et al.* 2017; Mo *et al.* 2019). Optical mapping of rDNA estimates that the genome of strain DMKU3-1042 contains at least 140 copies of rRNA (Lertwattanasakul *et al.* 2015).



**Figure 2.** Phylogenetic tree of yeasts including pre-WGD and post-WGD genera. This tree was adopted from Dashko *et al.* 2014. Blue arrows note two major evolution events including the whole genome duplication and the loss of respiratory complex I.

### ***Kluyveromyces marxianus* in food and industry**

For industrial applications, there is an interest in using *K. marxianus* as a cell factory to produce a wide range of valuable compounds (Rajaei *et al.* 2014; Morrissey *et al.* 2015; Varela *et al.* 2017; Karim, Gerliani and Aider 2020). On an industrial-scale, *K. marxianus* can sustain growth at temperatures exceeding 40°C, allowing industries to use a reduced cooling capacity as well as antimicrobial agents as many microbes will fail to grow at high temperatures. A rapid growth rate allows accumulation of biomass in a shorter time compared to other species. Studies have demonstrated the use of *K. marxianus* to express heterologous proteins which have commercial applications (reviewed in Gombert *et al.* 2016) such a beta-galactosidase (Yang *et al.* 2015) and glucoamylase (Raimondi *et al.* 2013). As carbon sources and other substrates required for growth can come at a significant financial cost to industries, using cheap by-products of other industries as a carbon source for *K. marxianus* has also been investigated. One example is whey, a by-product of cheese manufacturing which contains high concentrations of lactose, and it has been shown that *K. marxianus* is capable of using whey as a carbon source (Caballero *et al.* 1995). *K. marxianus* is capable of producing bioethanol while growing in presence of whey (Zafar and Owais 2006; Ozmihci and Kargi 2007). Other examples of bioethanol production on various substrates include corn silage juice (Hang, Woodams and Hang 2003), Jerusalem artichokes (Kim, Park and Kim 2013) and molasses (Martínez *et al.* 2017). Other cell factory applications include the production of the aromatic alcohol 2-Phenylethanol from genetically modified strains exploiting the Ehrlich pathway (Kim, Lee and Oh 2014; Rajkumar and Morrissey 2020), and glycerol production commonly used in industries as a solvent (Zhang *et al.* 2020).

*K. marxianus* also plays an important role in the food industry. Due to lactose assimilation properties, *K. marxianus* has commonly been associated with and isolated from dairy products including cheese, yoghurt and fermented milk beverages (Lachance 1998; Gethins *et al.* 2016; Coloretti *et al.* 2017). The association with food products has allowed this yeast to achieve GRAS (generally regarded as safe) and QPS (qualified presumption of safety) qualifications from the US and EU, respectively ((BIOHAZ) *et al.* 2022). This includes the *K. marxianus* B0399 strain which was isolated from milk and shown to survive the gastrointestinal tract and has been suggested as a suitable probiotic strain (Maccaferri *et al.* 2012; Tabanelli *et al.* 2016). This yeast has also been suggested a promising probiotic having

antioxidative, anti-inflammatory and cholesterol reducing properties (Xie *et al.* 2015; Cho *et al.* 2018).

### **Development of synthetic biology tools**

To play a role in any industrial setting, it must be possible to engineer or edit the genome of microbes to create efficient cell factories to synthesize valuable compounds. These include the addition of an exogenous gene(s) via introduction or a plasmid (engineering) or integration into the genome (editing). *In vivo* techniques of genome engineering/edit include the assembly of large gene products via the PGASO method (Chang *et al.* 2012) and more recently, the development of CRISPR/Cas9 system specific to *K. marxianus* (Löbs *et al.* 2017; Nambu-Nishida *et al.* 2017; Cernak *et al.* 2018; Juergens *et al.* 2018; Rajkumar *et al.* 2019). A popular method of cloning in *S. cerevisiae* includes the yeast toolkit (YTK), which allows *in vitro* assembly of vectors golden gate assembly method to assemble (in a single reaction) a vector from a list of characterized parts including promoters, terminators and several selection markers (Lee *et al.* 2015). From a library of these “parts” which includes promoters, terminators, selection markers and genes of interest, a single *in vitro* reaction utilizing DNA ligases can integrate all elements of the desired vector to express an exogenous gene(s). Recently, a YTK has been developed for *K. marxianus* and provides vector backbones for expressing heterologous proteins or CRISPR/Cas9 genome editing (Rajkumar *et al.* 2019).

### **Investigation of Global Responses to Stress**

A large effort has aimed at studying how *K. marxianus* responds to wide range of industrially relevant stresses and growth on various substrates. These studies are important to understand how *K. marxianus* can adapt to a specific condition but to also aid in generating targets for genome engineering and editing. These studies have included mostly RNA-seq experiments, using relative changes in mRNA abundances to determine gene expression changes under various industrially-relevant conditions including growth on inulin (Gao *et al.* 2015), high ethanol concentrations (6%) (Diniz *et al.* 2021), high temperatures (45°C) (Lertwattanasakul *et al.* 2015) and growth on compounds which are present in lignocellulosic substrates that inhibit growth, such as furfural, acetic acid and phenols (Wang *et al.* 2018). In one major study, *S. cerevisiae*, *Y. lipolytica* and *K. marxianus* were subjected to a range of stresses including low pH, high salt and high temperatures. Using a combination of RNA-seq and



proteomic mass spectrometry, this revealed that differentially expressed genes in these stressful conditions are enriched with evolutionary young genes (genes unique to each genus or species), which are suggested to play important roles in long-term adaptation to the stresses studied (Doughty *et al.* 2020). Later, using a CRISPR-cas9 knockout of these young genes in *K. marxianus*, it was discovered one such gene is vital for growth at high temperatures (Montini *et al.* 2022). These examples demonstrate how these yeasts have evolved to adapt to niche environments including high temperatures. These adaptations offer the potential for new industrial opportunities utilising unviable substrates or growth conditions which are often available at much lower costs or even as by-products to other industrial processes. Using RNA-seq and ribosome profiling, a full understanding of how these genes are regulated on both a transcriptional and a translational level under industrially relevant conditions could pave the way for newly engineered strains with a number of key benefits over current industrial strains. Newly identified genes and regulation has the opportunity to unlock new avenues of research into how environmental adaptations evolve, providing information on yeast gene expression regulation as well as offering cost effective solutions to current industrial pitfalls in using yeast as cell factories.

## References

- (BIOHAZ) EP on BH, Koutsoumanis K, Allende A *et al.* Update of the list of QPS-recommended biological agents intentionally added to food or feed as notified to EFSA 15: suitability of taxonomic units notified to EFSA until September 2021. *EFSA journal Eur Food Saf Auth* 2022;**20**:e07045–e07045.
- Binati RL, Salvetti E, Bzducha-Wróbel A *et al.* Non-conventional yeasts for food and additives production in a circular economy perspective. *FEMS Yeast Res* 2021;**21**:foab052.
- Botstein D, Fink GR. Yeast: an experimental organism for 21st Century biology. *Genetics* 2011;**189**:695–704.
- Caballero R, Olguín P, Cruz-Guerrero A *et al.* Evaluation of *Kluyveromyces marxianus* as baker's yeast. *Food Res Int* 1995;**28**:37–41.
- Carl M, Rosemary Y, Johan B *et al.* Adaptations in metabolism and protein translation give rise to the Crabtree effect in yeast. *Proc Natl Acad Sci* 2021;**118**:e2112836118.
- Cavalieri D, McGovern PE, Hartl DL *et al.* Evidence for *S. cerevisiae* Fermentation in Ancient Wine. *J Mol Evol* 2003;**57**:S226–32.
- Cernak P, Estrela R, Poddar S *et al.* Engineering *Kluyveromyces marxianus* as a Robust Synthetic Biology Platform Host. *MBio* 2018;**9**, DOI: 10.1128/mBio.01410-18.
- Chang J-J, Ho C-Y, Ho F-J *et al.* PGASO: A synthetic biology tool for engineering a cellulolytic yeast. *Biotechnol Biofuels* 2012;**5**:53.
- Cho Y-J, Kim D-H, Jeong D *et al.* Characterization of yeasts isolated from kefir as a probiotic and its synergic interaction with the wine byproduct grape seed flour/extract. *LWT* 2018;**90**:535–9.
- Coloretti F, Chiavari C, Luise D *et al.* Detection and identification of yeasts in natural whey starter for Parmigiano Reggiano cheese-making. *Int Dairy J* 2017;**66**:13–7.
- Dasgupta D, Ghosh D, Bandhu S *et al.* Lignocellulosic sugar management for xylitol and ethanol fermentation with multiple cell recycling by *Kluyveromyces marxianus* IIPE453. *Microbiol Res* 2017;**200**:64–72.
- Dashko S, Zhou N, Compagno C *et al.* Why, when and how did yeast evolve alcoholic fermentation? *FEMS Yeast Res* 2014;**14**, DOI: 10.1111/1567-1364.12161.
- Desfougères T, Haddouche R, Fudalej F *et al.* SOA genes encode proteins controlling lipase expression in response to triacylglycerol utilization in the yeast *Yarrowia lipolytica*. *FEMS Yeast Res* 2010;**10**:93–103.

- Diniz RHS, Villada JC, Alvim MCT *et al.* Transcriptome analysis of the thermotolerant yeast *Kluyveromyces marxianus* CCT 7735 under ethanol stress. *Appl Microbiol Biotechnol* 2021;**105**:2613.
- Donzella L, Varela JA, Sousa MJ *et al.* Identification of novel pentose transporters in *Kluyveromyces marxianus* using a new screening platform. *FEMS Yeast Res* 2021;**21**, DOI: 10.1093/femsyr/foab026.
- Doughty TW, Domenzain I, Millan-Oropeza A *et al.* Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts. *Nat Commun* 2020;**11**:2144.
- Duman-Özdamar ZE, Binay B. Production of Industrial Enzymes via *Pichia pastoris* as a Cell Factory in Bioreactor: Current Status and Future Aspects. *Protein J* 2021;**40**:367–76.
- Fonseca GG, Heinzle E, Wittmann C *et al.* The yeast *Kluyveromyces marxianus* and its biotechnological potential. *Appl Microbiol Biotechnol* 2008;**79**:339–54.
- Gao J, Yuan W, Li Y *et al.* Transcriptional analysis of *Kluyveromyces marxianus* for ethanol production from inulin using consolidated bioprocessing technology. *Biotechnol Biofuels* 2015;**8**:115.
- Gethins L, Rea MC, Stanton C *et al.* Acquisition of the yeast *Kluyveromyces marxianus* from unpasteurised milk by a kefir grain enhances kefir quality. *FEMS Microbiol Lett* 2016;**363**, DOI: 10.1093/femsle/fnw165.
- Goffeau A, Barrell BG, Bussey H *et al.* Life with 6000 genes. *Science* 1996;**274**:546,563–567.
- Gombert AK, Madeira JVJ, Cerdán M-E *et al.* *Kluyveromyces marxianus* as a host for heterologous protein synthesis. *Appl Microbiol Biotechnol* 2016;**100**:6193–208.
- Gonçalves FAG, Colen G, Takahashi JA. *Yarrowia lipolytica* and Its Multiple Applications in the Biotechnological Industry. Dumais N, Heng N, Yan Y (eds.). *Sci World J* 2014;**2014**:476207.
- Groeneveld P, Stouthamer AH, Westerhoff H V. Super life--how and why “cell selection” leads to the fastest-growing eukaryote. *FEBS J* 2009;**276**:254–70.
- Hang YD, Woodams EE, Hang LE. Utilization of corn silage juice by *Kluyveromyces marxianus*. *Bioresour Technol* 2003;**86**:305–7.
- Hittinger CT, Rokas A, Bai F-Y *et al.* Genomics and the making of yeast biodiversity. *Curr Opin Genet Dev* 2015;**35**:100–9.
- Inokuma K, Ishii J, Hara KY *et al.* Complete Genome Sequence of *Kluyveromyces marxianus* NBRC1777, a Nonconventional Thermotolerant Yeast. *Genome Announc*

- 2015;**3**:e00389-15.
- Jeong H, Lee D-H, Kim SH *et al.* Genome sequence of the thermotolerant yeast *Kluyveromyces marxianus* var. *marxianus* KCTC 17555. *Eukaryot Cell* 2012;**11**:1584–5.
- Juergens H, Varela JA, Gorter de Vries AR *et al.* Genome editing in *Kluyveromyces* and *Ogataea* yeasts using a broad-host-range Cas9/gRNA co-expression plasmid. *FEMS Yeast Res* 2018;**18**:foy012.
- Karim A, Gerliani N, Aïder M. *Kluyveromyces marxianus*: An emerging yeast cell factory for applications in food and biotechnology. *Int J Food Microbiol* 2020;**333**:108818.
- Kim S, Park JM, Kim CH. Ethanol Production Using Whole Plant Biomass of Jerusalem Artichoke by *Kluyveromyces marxianus* CBS1555. *Appl Biochem Biotechnol* 2013;**169**:1531–45.
- Kim T-Y, Lee S-W, Oh M-K. Biosynthesis of 2-phenylethanol from glucose with genetically engineered *Kluyveromyces marxianus*. *Enzyme Microb Technol* 2014;**61–62**:44–7.
- Krause DJ, Kominek J, Opulente DA *et al.* Functional and evolutionary characterization of a secondary metabolite gene cluster in budding yeasts. *Proc Natl Acad Sci U S A* 2018;**115**:11030–5.
- Lachance MA. 36 - *Kluyveromyces van der Walt* emend. *van der Walt*. In: Kurtzman CP, Fell JWBT-TY (Fourth E (eds.). Amsterdam: Elsevier, 1998, 227–47.
- Lane MM, Burke N, Karreman R *et al.* Physiological and metabolic diversity in the yeast *Kluyveromyces marxianus*. *Antonie Van Leeuwenhoek* 2011;**100**:507–19.
- Lane MM, Morrissey JP. *Kluyveromyces marxianus*: A yeast emerging from its sister's shadow. *Fungal Biol Rev* 2010;**24**:17–26.
- Lee ME, DeLoache WC, Cervantes B *et al.* A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. *ACS Synth Biol* 2015;**4**:975–86.
- Lertwattanasakul N, Kosaka T, Hosoyama A *et al.* Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. *Biotechnol Biofuels* 2015;**8**:47.
- Löbs A-K, Engel R, Schwartz C *et al.* CRISPR–Cas9-enabled genetic disruptions for understanding ethanol and ethyl acetate biosynthesis in *Kluyveromyces marxianus*. *Biotechnol Biofuels* 2017;**10**:164.
- Maccaferri S, Klinder A, Brigidi P *et al.* Potential probiotic *Kluyveromyces marxianus* B0399 modulates the immune response in Caco-2 cells and peripheral blood

- mononuclear cells and impacts the human gut microbiota in an in vitro colonic model system. *Appl Environ Microbiol* 2012;**78**:956–64.
- Martínez O, Sánchez A, Font X *et al.* Valorization of sugarcane bagasse and sugar beet molasses using *Kluyveromyces marxianus* for producing value-added aroma compounds via solid-state fermentation. *J Clean Prod* 2017;**158**:8–17.
- Mo W, Wang M, Zhan R *et al.* *Kluyveromyces marxianus* developing ethanol tolerance during adaptive evolution with significant improvements of multiple pathways. *Biotechnol Biofuels* 2019;**12**:63.
- Montini N, Doughty TW, Domenzain I *et al.* Identification of a novel gene required for competitive growth at high temperature in the thermotolerant yeast *Kluyveromyces marxianus*. *Microbiology* 2022;**168**, DOI: 10.1099/mic.0.001148.
- Morrissey JP, Etschmann MMW, Schrader J *et al.* Cell factory applications of the yeast *Kluyveromyces marxianus* for the biotechnological production of natural flavour and fragrance molecules. *Yeast* 2015;**32**:3–16.
- Nambu-Nishida Y, Nishida K, Hasunuma T *et al.* Development of a comprehensive set of tools for genome engineering in a cold- and thermo-tolerant *Kluyveromyces marxianus* yeast strain. *Sci Rep* 2017;**7**:8993.
- Nandy SK, Srivastava RK. A review on sustainable yeast biotechnological processes and applications. *Microbiol Res* 2018;**207**:83–90.
- Nielsen J. Yeast Systems Biology: Model Organism and Cell Factory. *Biotechnol J* 2019;**14**:1800421.
- Ozmihci S, Kargi F. Kinetics of batch ethanol fermentation of cheese-whey powder (CWP) solution as function of substrate and yeast concentrations. *Bioresour Technol* 2007;**98**:2978–84.
- Parapouli M, Vasileiadis A, Afendra A-S *et al.* *Saccharomyces cerevisiae* and its industrial applications. *AIMS Microbiol* 2020;**6**:1–31.
- Raimondi S, Zanni E, Amaretti A *et al.* Thermal adaptability of *Kluyveromyces marxianus* in recombinant protein production. *Microb Cell Fact* 2013;**12**:34.
- Rajaei N, Chiruvella KK, Lin F *et al.* Domesticated transposase Kat1 and its fossil imprints induce sexual differentiation in yeast. *Proc Natl Acad Sci* 2014;**111**:15491 LP – 15496.
- Rajkumar AS, Morrissey JP. Rational engineering of *Kluyveromyces marxianus* to create a chassis for the production of aromatic products. *Microb Cell Fact* 2020;**19**:207.
- Rajkumar AS, Varela JA, Juergens H *et al.* Biological Parts for *Kluyveromyces marxianus*

- Synthetic Biology . *Front Bioeng Biotechnol* 2019;**7**:97.
- Rebello S, Abraham A, Madhavan A *et al*. Non-conventional yeast cell factories for sustainable bioprocesses. *FEMS Microbiol Lett* 2018;**365**:fny222.
- Schaffrath R, Breunig KD. Genetics and molecular physiology of the yeast *Kluyveromyces lactis*. *Fungal Genet Biol* 2000;**30**:173–90.
- Shen X-X, Opulente DA, Kominek J *et al*. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* 2018;**175**:1533-1545.e20.
- Shevchenko A, Yang Y, Knaust A *et al*. Proteomics identifies the composition and manufacturing recipe of the 2500-year old sourdough bread from Subeixi cemetery in China. *J Proteomics* 2014;**105**:363–71.
- Sicard D, Legras J-L. Bread, beer and wine: yeast domestication in the *Saccharomyces sensu stricto* complex. *C R Biol* 2011;**334**:229–36.
- Suzuki T, Hoshino T, Matsushika A. Draft Genome Sequence of *Kluyveromyces marxianus* Strain DMB1, Isolated from Sugarcane Bagasse Hydrolysate. *Genome Announc* 2014;**2**, DOI: 10.1128/genomeA.00733-14.
- Tabanelli G, Verardo V, Pasini F *et al*. Survival of the functional yeast *Kluyveromyces marxianus* B0399 in fermented milk with added sorbic acid. *J Dairy Sci* 2016;**99**:120–9.
- Varela JA, Gethins L, Stanton C *et al*. Applications of *Kluyveromyces marxianus* in Biotechnology BT - Yeast Diversity in Human Welfare. In: Satyanarayana T, Kunze G (eds.). Singapore: Springer Singapore, 2017, 439–53.
- Wang D, Wu D, Yang X *et al*. Transcriptomic analysis of thermotolerant yeast *Kluyveromyces marxianus* in multiple inhibitors tolerance. *RSC Adv* 2018;**8**:14177–92.
- Winzeler EA, Shoemaker DD, Astromoff A *et al*. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 1999;**285**:901–6.
- Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 1997;**387**:708–13.
- Xie Y, Zhang H, Liu H *et al*. Hypocholesterolemic effects of *Kluyveromyces marxianus* M3 isolated from Tibetan mushrooms on diet-induced hypercholesterolemia in rat. *Braz J Microbiol* 2015;**46**:389–95.
- Yamada R. Chapter 17 - *Pichia pastoris*-based microbial cell factories. In: Singh VBT-MCFE for P of B (ed.). Academic Press, 2021, 335–44.
- Yang C, Hu S, Zhu S *et al*. Characterizing yeast promoters used in *Kluyveromyces marxianus*. *World J Microbiol Biotechnol* 2015;**31**:1641–6.

- Zafar S, Owais M. Ethanol production from crude whey by *Kluyveromyces marxianus*. *Biochem Eng J* 2006;**27**:295–8.
- Zhang B, Ren L, Wang Y *et al*. Glycerol production through TPI1 defective *Kluyveromyces marxianus* at high temperature with glucose, fructose, and xylose as feedstock. *Biochem Eng J* 2020;**161**:107689.
- Zhu T, Sun H, Wang M *et al*. *Pichia pastoris* as a Versatile Cell Factory for the Production of Industrial Enzymes and Chemicals: Current Status and Future Perspectives. *Biotechnol J* 2019;**14**:1800694.

## Aims and Objectives of this thesis

*Kluyveromyces marxianus* is a thermotolerant yeast with a broad range of traits which make this yeast attractive for biotechnology applications. While many studies focus on transcriptomic techniques to study changes in mRNA abundance, changes in translation of these mRNAs therefore ignored. Ribosome profiling is a powerful tool to study mRNA translation in the cell, revealing the locations of translating ribosomes on mRNAs. In this thesis, the initial goal is the development and demonstration of a working ribosome profiling protocol for *Kluyveromyces marxianus*.

Once a ribosome profiling protocol for *K. marxianus* is established, ribosome profiling data is analysed alongside transcriptomic techniques to study the translatome and transcriptome of *K. marxianus*. These data reveal previously unannotated genes, mechanisms of mRNA translation leading to generation of alternative proteoforms and potential regulation.

As industrial yeasts are subjected to various stresses in industrial settings, this thesis employs ribosome profiling to study how *K. marxianus* adapts to high temperatures. This study relies on a timecourse to study both early and late adaptations to an increase in temperature.

Finally, this thesis will discuss how ribosome profiling may be a useful tool to study industrially relevant stresses in other NCYs.



## Chapter 2 - Development of a Ribosome Profiling Protocol to Study Translation in *Kluyveromyces marxianus*

*This chapter has been published as a paper in FEMS Yeast Research 2022*

<https://academic.oup.com/femsyr/advance-article/doi/10.1093/femsyr/foac024/6581590>

### Abstract

*Kluyveromyces marxianus* is an interesting and important yeast because of particular traits like thermotolerance and rapid growth, and applications in food and industrial biotechnology. Both for understanding its biology and for developing bioprocesses, it is important to understand how *K. marxianus* responds and adapts to changing environments. For this, a full suite of omics tools to measure and compare global patterns of gene expression and protein synthesis is needed. We report here the development of a ribosome profiling method for *K. marxianus*, which allows codon-resolution of translation on a genome-wide scale by deep sequencing of ribosome locations on mRNAs. To aid in the analysis and sharing of ribosome profiling data, we also added the *K. marxianus* genome as well as transcriptome and ribosome profiling data to the publicly accessible GWIPS-viz and Trips-Viz browsers. Users are able to upload custom ribosome profiling and RNA-Seq data to both browsers, therefore allowing easy analysis and sharing of data. We also provide a set of step by step protocols for the experimental and bioinformatic methods that we developed.

## Introduction

As with other microbes, yeasts have evolved elaborate mechanisms to sense and respond to changing extracellular and intracellular environments. External influences include phenomena such as altered nutrient availability, toxic molecules, temperature fluctuations, low pH and high osmotic pressure, while internally, cells can experience changes such as reduced intracellular pH, ion fluxes, energy depletion or nutrient starvation (Martínez-Montañés, Pascual-Ahuir and Proft 2010; Broach 2012; Ljungdahl and Daignan-Fornier 2012; Morano, Grant and Moye-Rowley 2012; de la Torre-Ruiz, Pujol and Sundaran 2015; Sui *et al.* 2015; Taymaz-Nikerel, Cankorur-Cetinkaya and Kirdar 2016). The best-studied response mechanisms in yeast involve sensor systems, signal transduction pathways, and changes in gene expression (de Nadal and Posas 2010). Ultimately, this gives rise to a new set of proteins that enable the cell to adapt, if necessary, and to survive as well as proliferate in this new environment. Dissecting these response mechanisms is central to understanding the fundamental biology of a species, but it is also important for the development of yeast for biotechnological applications (Liu and Nielsen 2019). Yeasts are used for diverse applications in the food, biopharma and industrial biotechnology sectors (Arevalo-Villena *et al.* 2017; Nandy and Srivastava 2018; Parapouli *et al.* 2020) and, very often, they need to perform under suboptimal conditions or deal with a fluctuating environment. This is a particular problem when scaling engineered yeast cell factories in industrial biotechnology (Takors 2012; Delvigne *et al.* 2014; Wehrs *et al.* 2019). Developing a comprehensive understanding of adaptive responses is a key requirement for the construction of yeast cell factories that are both robust and resilient, and capable of optimal performance in an industrial bioprocess.

Most adaptive responses involve increased or reduced activity of specific proteins, which can be achieved at the level of synthesis, stability or activity. While some specific responses can be at the protein level, for example mediated by allosteric regulation, adaptation usually requires changes in the expression of many genes, and is considered to be a “global” response. Changes in global gene expression can occur at various levels, most notably, via transcription, translation or mRNA stability. Transcriptional changes are due to chromatin restructuring or changes in the activity of particular transcriptional regulators leading to increased or decreased levels of mRNA (Hahn and Young 2011). This is by far the best-understood adaptive process, deployed in response to heat shock (Masser, Ciccarelli and

Andréasson 2020), osmotic stress (de Nadal and Posas 2010), oxidative stress (Morano, Grant and Moye-Rowley 2012) and cell wall challenges (Sanz *et al.* 2017; Jiménez-Gutiérrez *et al.* 2020). Transcriptional responses can be studied at the level of individual genes using Northern blots and RT-qPCR, or globally by DNA microarrays or massively parallel sequencing (RNA-Seq), the latter of which has become the method of choice to study changes in gene expression (Schena *et al.* 1995; Gibson, Heid and Williams 1996; Wang, Gerstein and Snyder 2009). Translation results in protein synthesis and, as such, is a better indicator of protein levels than transcription though in many cases, higher abundance of mRNA due to increased transcription leads to a corresponding increase in the amount of translation. This is not always the case, however, and there are also instances where translation is regulated without any changes in the mRNA abundance. A well-documented example of this in yeast is regulation of the translation of the transcriptional activator encoded by *GCN4*. In this case, short open reading frames upstream of the main *GCN4* coding sequence regulate the rate of *GCN4* translation in response to intracellular amino acid levels (Hinnebusch 2005). The TOR growth control system also mediates some of its effects by regulating translation via controlling access of the small ribosomal subunit to the cap structure at the 5' end of the mRNA (Merrick 2015). Indeed, as will be mentioned below, there is increased awareness that translational regulation is a central part of the yeast system for controlling gene expression.

Ribosome profiling, sometimes termed Ribo-Seq, is a method that allows for the visualisation and quantification of translation at a global level. First developed in *S. cerevisiae* (Ingolia *et al.*, 2009), it has since been widely used in bacteria, yeast and mammalian systems for genome wide studies of translation (Andreev *et al.* 2017; Ingolia, Hussmann and Weissman 2019; Mohammad, Green and Buskirk 2019). During mRNA translation, a ribosome translocates an mRNA one codon at a time and protects a fragment of mRNA within its mRNA tunnel (Steitz 1969). Ribosome profiling is a method to identify these ribosome protected fragments (RPFs), thereby reporting what mRNAs are being translated at a given point in time. When applying the method, translation is arrested, usually by the addition of translation inhibitors, ribosomes are isolated, and the RPFs identified by deep sequencing. RNA-Seq is usually carried out in parallel to ribosome profiling, allowing estimation of changes in mRNA translation efficiency (Ingolia *et al.* 2009). In *S. cerevisiae*, ribosome profiling has uncovered wide-spread translation of upstream open reading frames (uORFs)

(Ingolia *et al.* 2009), non-AUG initiation at canonical genes (Monteuuis *et al.* 2019; Eisenberg *et al.* 2020) and small translated ORFs throughout the genome (Smith *et al.* 2014). Ribosome profiling combined with RNA-Seq has been useful deciphering both transcriptional and translation regulation in the yeast meiotic programme (Brar and Weissman 2015) and in the response to oxidative stress (Blevins *et al.* 2019). Ribosome profiling has also been carried out on a range of other yeast species including *Saccharomyces paradoxus* (McManus *et al.* 2014), *Schizosaccharomyces pombe* (Duncan and Mata 2014), *Saccharomyces uvarum* (Spealman *et al.* 2018), *Komagataella phaffii* (Alva, Riera and Chartron 2021) and *Candida albicans* (Sharma *et al.* 2021).

We are especially interested in another budding yeast, *Kluyveromyces marxianus*, which originally attracted interest because of its role in food fermentations (Coloretti *et al.* 2017) but is now increasingly being considered as a platform of industrial biotechnology (Fonseca *et al.* 2008; Lane and Morrissey 2010; Karim, Gerliani and Aider 2020). *K. marxianus* has some intrinsic traits like thermotolerance, a broad substrate range and rapid growth that are useful for biotechnology (Groeneveld, Stouthamer and Westerhoff 2009), and molecular and genomic tools to aid its development as an industrial platform (Cernak *et al.* 2018; Rajkumar *et al.* 2019; Rajkumar and Morrissey 2020). To date, all studies that addressed gene expression in this yeast focused on transcriptional effects via RNA-Seq experiments. Aspects that have been studied include growth and ethanol production on alternative sugar substrates such as xylose (Schabert *et al.* 2016; Kwon *et al.* 2019) and inulin (Gao *et al.* 2015); ethanol tolerance during adaptive laboratory evolution (Mo *et al.* 2019); response to growth inhibitors derived from lignocellulosic substrates (Wang *et al.* 2018) and the ability to grow at high temperatures (Fu *et al.* 2019). Recently, mainly using transcriptome analysis, we determined that young genes specific to *K. marxianus* are enriched in the response to stresses such as high temperature, low pH and high osmolarity (Doughty *et al.* 2020).

To complement the molecular toolbox, and as a resource to study the biology of this yeast, here we report the development of a protocol to carry out ribosome profiling in *K. marxianus*. For this, we adapted and applied the methods previously used for *S. cerevisiae* (Ingolia *et al.* 2009). We also developed a suite of bioinformatics tools to visualise and analyse *K. marxianus* RNA-Seq and ribosome profiling results. This involved addition of the *K. marxianus* data to publicly available genome (GWIPS-viz) and transcriptome (Trips-Viz)

browsers, which, in turn, can be uploaded with user-generated expression data and used in private or public configurations. To facilitate the use of ribosome profiling as a very valuable tool to explore gene expression, we also include a detailed step by step protocol for users.

## **Materials and Methods**

### Strains and growth conditions

*K. marxianus* strain CBS 6556 (CBS-KNAW culture collection, Westerdijk Institute) was used in these studies following standard growth and handling procedures. This particular strain is also available from other collections under the strain name ATCC 26548, NRRL Y-7571, KCTC 17555 and NCYC 2597 and has been quite widely used as a representative *K. marxianus* strain. For ribosome profiling experiments, standard growth conditions used synthetic minimal medium (Verduyn *et al.* 1992) and an incubation temperature of 30°C with shaking. The mineral medium consisted of the following per litre amounts: (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 5.0 g; KH<sub>2</sub>PO<sub>4</sub>, 3.0 g; MgSO<sub>4</sub>·7H<sub>2</sub>O, 0.5 g; trace elements (EDTA, 15 mg; ZnSO<sub>4</sub>·7H<sub>2</sub>O, 4.5 mg; MnCl<sub>2</sub>·2H<sub>2</sub>O, 0.84 mg; CoCl<sub>2</sub>·6H<sub>2</sub>O, 0.3 mg; CuSO<sub>4</sub>·5H<sub>2</sub>O, 0.3 ; Na<sub>2</sub>MoO<sub>4</sub>·2H<sub>2</sub>O, 0.4 ; CaCl<sub>2</sub>·2H<sub>2</sub>O, 4.5 mg; FeSO<sub>4</sub>·7H<sub>2</sub>O, 3.0 mg; H<sub>3</sub>BO<sub>3</sub>, 1.0 mg; KI, 0.1 mg); silicone antifoam, 0.05 mL. The medium was adjusted to pH 6.0 with KOH before autoclaving (121°C, 20 min). The medium was cooled to room temperature and a filter-sterilized solution of vitamins prepared in demineralized water was added, to a final concentration, per liter, of: d-biotin, 0.05 mg; calcium pantothenate, 1.0 mg; nicotinic acid, 1.0 mg; *myo*-inositol, 25 mg; thiamine HCl, 1.0 mg; pyridoxin HCl, 1.0 mg; and para-aminobenzoic acid, 0.20 mg. Glucose was sterilized separately and added to a final concentration of 10 g L<sup>-1</sup>. 150 mL cultures in 500 mL conical flasks were grown to early-log phase at A<sub>600</sub> ~0.8 and either harvested or transferred to a shaking water bath at 40°C, with cells harvested at 5, 15, 30 and 60 minutes. All experiments were carried out with two biological replicates.

### Ribosome Profiling

For cell harvesting, cultures were quickly poured into a glass filter assembly (Durapore) with using 0.45 µm pore nitrocellulose filter membrane (GE #7184-009). A vacuum pump was immediately turned on and once liquid media was removed, cells were quickly scraped into a 50 mL falcon tube filled with liquid nitrogen. Once a 150 mL culture is added to the filtration assembly, it takes ~9-12 seconds until the media is removed and the scraped cell pellet is collected and submerged in liquid nitrogen. After harvesting, 1.5 mL of polysome lysis buffer

(5 mM MgCl<sub>2</sub>, 150 mM KCl, 20 mM Tris-HCl, 100 µg/mL cycloheximide, 1 mM DTT, 1% Triton X-100) was slowly added dropwise to the liquid nitrogen and cells to create a frozen mixture of buffer and cells. The 50 mL tube (with pierced cap from screwdriver) was placed in -80°C to allow boiling off of the liquid nitrogen. Frozen cells/buffer were disrupted using cryogenic grinding using a Retsch Mixer Mill 400 and 10 mL steel grinding jars and balls. Samples were ground for 6 cycles of 3 minutes each at 20 Hz, the steel jars were submerged in liquid nitrogen to cool samples between each cycle. After lysis, lysates were gently thawed on ice and quantified with Qubit 4.0 fluorometer and BR-assay kit (#Q10211, Invitrogen). 30 µg of total RNA from the lysate was diluted to 200 µL in polysome buffer (lysis buffer without Triton X-100) and 1.5 µL RNase I was added (Epicentre #N6901K). RNase digestion was carried out at 200 rpm at 37°C for 45 minutes. To halt digestion, SUPERase•In (Invitrogen) was added, and samples were placed on ice before loading onto cold 10-50% sucrose gradients, which were prepared using a Biocomp Gradient Master. Gradients were spun for 3 hours at 4°C and 36,000 RPM (221,632 x g) on SW41-Ti rotor (Beckman Coulter). Monosome fractions were isolated from each sucrose gradient with Brandel Density Gradient Fractionator using 1.5 mL/min flow speed and 60% CsCl, aliquoting fractions every 12 seconds on a UV-visible 96 well plate. Reading the 96 well plate at 260 nm determined which well(s) contained the monosome fractions. RNA from monosome fractions were isolated using Trizol (Invitrogen) (Chomczynski and Sacchi 2006). Ribosome footprints were size selected with a 15% PAGE-Urea gel (70 minutes in 1X TBE and 300 V constant) using a 26 and 34 nt RNA marker (IDT) as a guide for excision. Using a scalpel, a slice representing the RPFs was cut from gel and placed into a 1.5 mL RNase-free Eppendorf tube and 500 µL of RNA elution buffer (300 mM NaOAc pH 5.5, 1 mM EDTA and 0.25% v/v SDS) was added. Following overnight shaking at room temperature to elute the RPFs, RPFs were precipitated using standard alcohol precipitation using ice-cold isopropanol, 80% ethanol and 1.5 µL Glycoblue co-precipitant (Ambion #AM9515).

### Library Construction

cDNA library construction with ribosome footprints is based on McGlincy et al. 2017 (McGlincy and Ingolia 2017) protocol with minor modifications. In brief, size selected ribosome footprints were treated with T4 Polynucleotide Kinase (#M0201L, New England Biolabs (NEB)) followed by ligation to a DNA linker using T4 RNA ligase 2, truncated K227Q (#M0351L, NEB). The footprints were reverse transcribed using Protoscript II (#M0368L, NEB). cDNA products were circularized using circligase II (#CL9025K,

Epicentre). The major rRNA contaminants were removed using subtractive hybridization with custom biotinylated oligos (Sigma Aldrich) and streptavidin beads (#65001, Invitrogen) as described in Ingolia *et al.*, 2012. The remaining circularized products were amplified by PCR using Phusion polymerase (#M0530L, NEB). In a pilot experiments, libraries were sequenced on MiSeq platform at the Teagasc Next Generation DNA Sequencing Facility, Moorepark, Moorepark West, Fermoy, Co. Cork, Ireland. Prepared libraries using the protocol described within were sequenced on Illumina HiSeq 4000 using SE-75 sequencing at the Genomics & Cell Characterization Core Facility (GC3F), University of Oregon, Eugene, Oregon, USA.

#### Ribosome profiling data analysis pre-processing & genome annotation update

Adapter sequences were removed from reads using Cutadapt (Martin 2011). For genomic alignments, rRNA contaminants were removed and remaining reads were aligned to the *K. marxianus* DMKU3-1042 reference genome (Lertwattanasakul *et al.* 2015) with Bowtie (Langmead *et al.* 2009) using parameters -n 2 -m 1 (these parameters allow 2 nt mismatches and filter out reads aligning to two or more locations). Transcriptome alignments were made with Trips-Viz (Kiniry *et al.* 2019). For all genomic and transcriptome analysis, Ribosome profiling and RNA-Seq reads were aligned to the *K. marxianus* DMKU3-1042 reference strain (Lertwattanasakul *et al.* 2015).

#### RNA-Seq

RNA was isolated from clarified lysates using Trizol (Chomczynski and Sacchi 2006) (Invitrogen #15596026) and quantified with a Qubit 4.0 fluorometer (Invitrogen). 1 µg of total RNA from each sample was analysed on an agarose bleach gel to determine RNA quality. Samples were sent to BGI Hong Kong for yeast rRNA removal (Ribo-Zero Gold rRNA Removal kit by Illumina (now discontinued)), library generation and sequencing with paired-end chemistry. Alternatively, RNA-Seq was carried out using polyA selection using Poly(A)Purist Mag Kit (Ambion #AM1922) as per manufacturer's instructions. PolyA selected cDNA library was generated and sequenced in the same method as ribosome profiling.

## Results and Discussion

### Development of a Ribosome Profiling Protocol to study translation in *K. marxianus*

The previously-described *S. cerevisiae* protocol (McGlinchy and Ingolia 2017) was used as the basis for development of a ribosome profiling procedure for *K. marxianus*. An overview of the pipeline from culturing to downstream bioinformatic analyses is shown in **Fig. 1**.

Summarising the first part of the protocol, cultures are rapidly harvested and flash frozen to preserve the translational state of the cell. Frozen cells are then lysed cryogenically in the presence of cycloheximide, which ensures that ribosomes remain stalled even if the cells thaw, and the clarified lysate containing polysomes is treated with RNase I to digest unprotected mRNA surrounding the ribosomes, retaining the ribosome protected fragment (RPF). Monosomes are isolated from a sucrose gradient and loaded on a polyacrylamide gel to allow size selection of RPFs of ~28nt. A cDNA library of these fragments is created and sequenced to identify the RPFs.

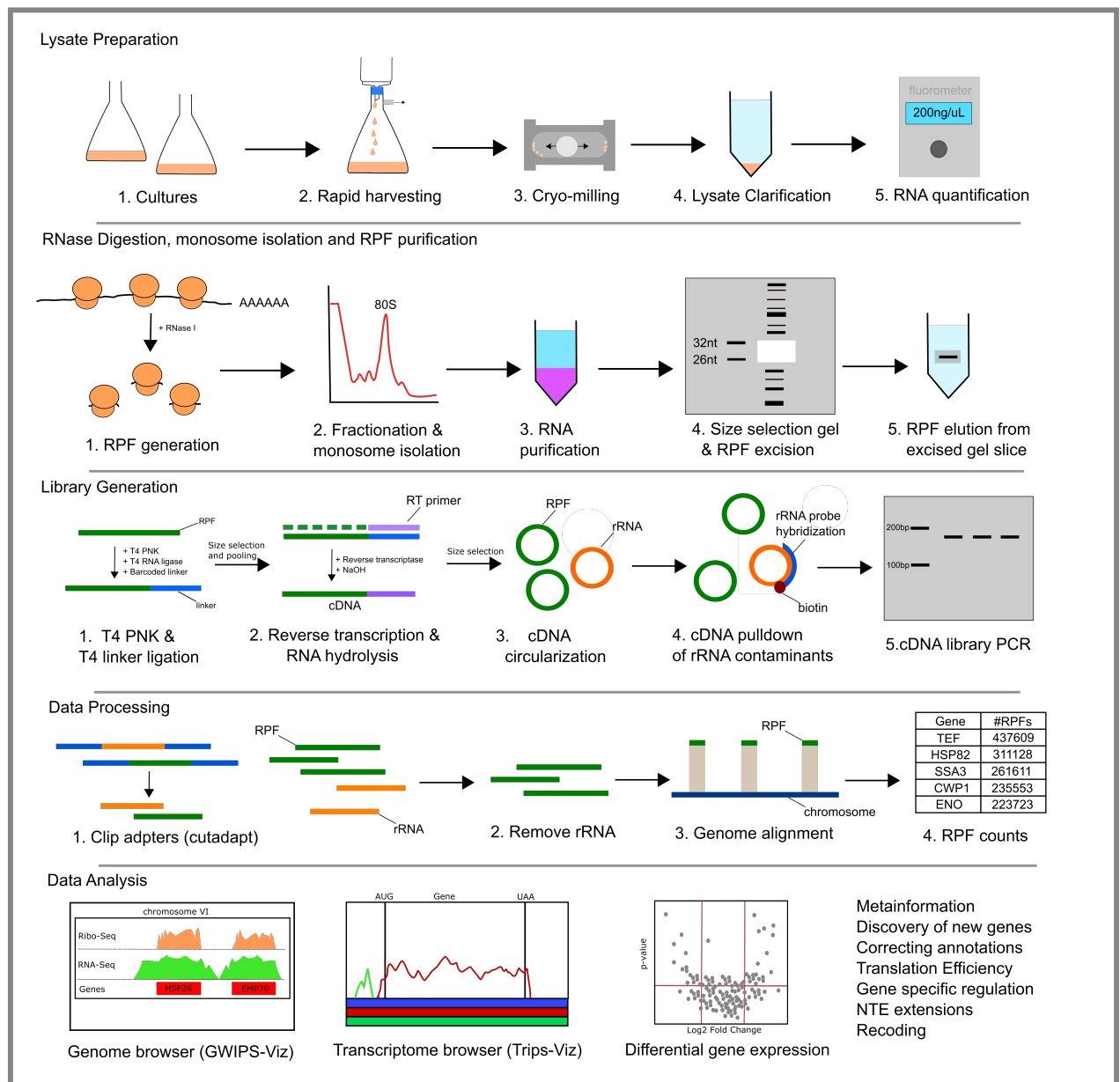
A limited-scale pilot experiment was first carried out to validate the methods and to identify the most abundant rRNA contaminants in the library. These arise because of RNase I digestion of rRNA and subsequent co-purification of fragments of the same size as the RPFs. Due to natural polymorphisms in the rRNA encoding genes between species, the sequence of the major contaminated rRNA fragments needs to be determined empirically for each yeast. Knowing these sequences allows the design of synthetic biotinylated oligos that can be used to reduce rRNA contamination (Ingolia *et al.* 2012). In our pilot library, we identified six highly abundant rRNA fragments, four from the 25S rRNA, and one each from the 18S rRNA and from the mitochondrial 21S rRNA (**Table 1**).

name	rRNA target	Sequence
rRNA#1	26S	5'AAGGGTGCATCATCGACCGATCCTG 3'
rRNA#2	26S	5'GTTTCTTTACTTATTCAATTAAGCGGA 3'
rRNA#3	mitochondrial	5'TAAAGAATGGTACAGCTATAAATATT 3'
rRNA#4	18S	5'GCTCGAATATATTAGCATGGAATAATGGA 3'
rRNA#5	26S	5'TATAGAAGGATACGAATAAGGCGTC 3'
rRNA#6	26S	5'TTTCCACGTTCTAGCATTCAAAGTCCT 3'

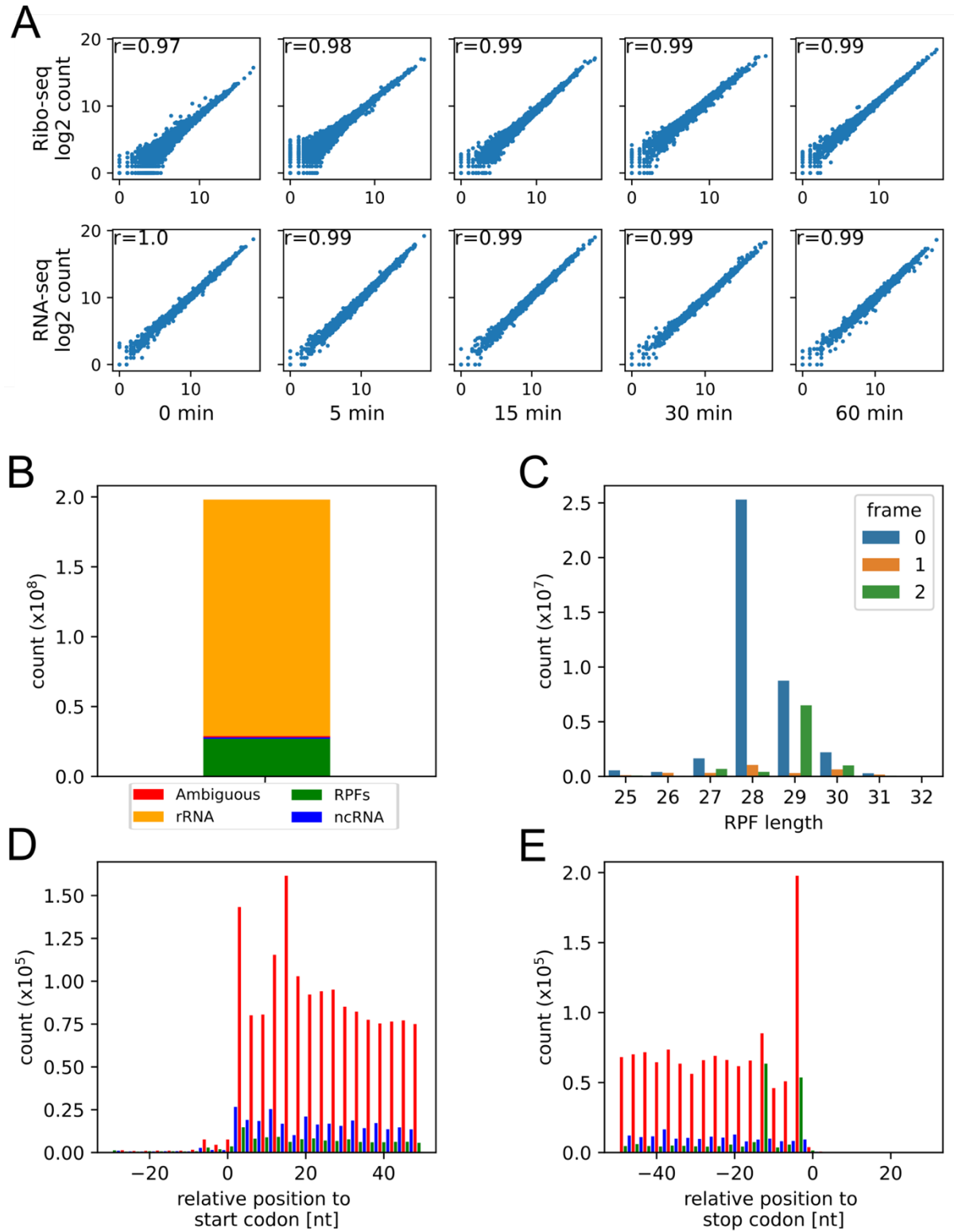
**Table 1.** Biotinylated oligos for rRNA depletion. These oligos contain a 5' biotin modification to allow pulldown of specific rRNA contaminants using magnetic streptavidin beads.



One of these sequences from 25S rRNA (GGGTGCATCATCGACCGATCCT) comprised ~33% of all rRNA contaminants. By reducing rRNA contamination, the proportion of RPFs in a library is increased and thus more usable data are generated per experiment. The ribosome profiling protocol with rRNA depletion was then tested on a larger scale using 150 mL cultures of *K. marxianus* growing at different temperatures to increase the total number of genes that would be expressed. The focus at this time was on assessing the quality of the data generated and the robustness of the protocol rather than on analysis of changes in gene expression. Ribosome profiling was performed in duplicate on flask cultures at 30°C and at 5, 15, 30 and 60 minutes after a transfer from 30°C to 40°C and, as is standard, RNA-Seq was also performed to measure transcript levels. Several analyses were performed to assess the robustness of the data that



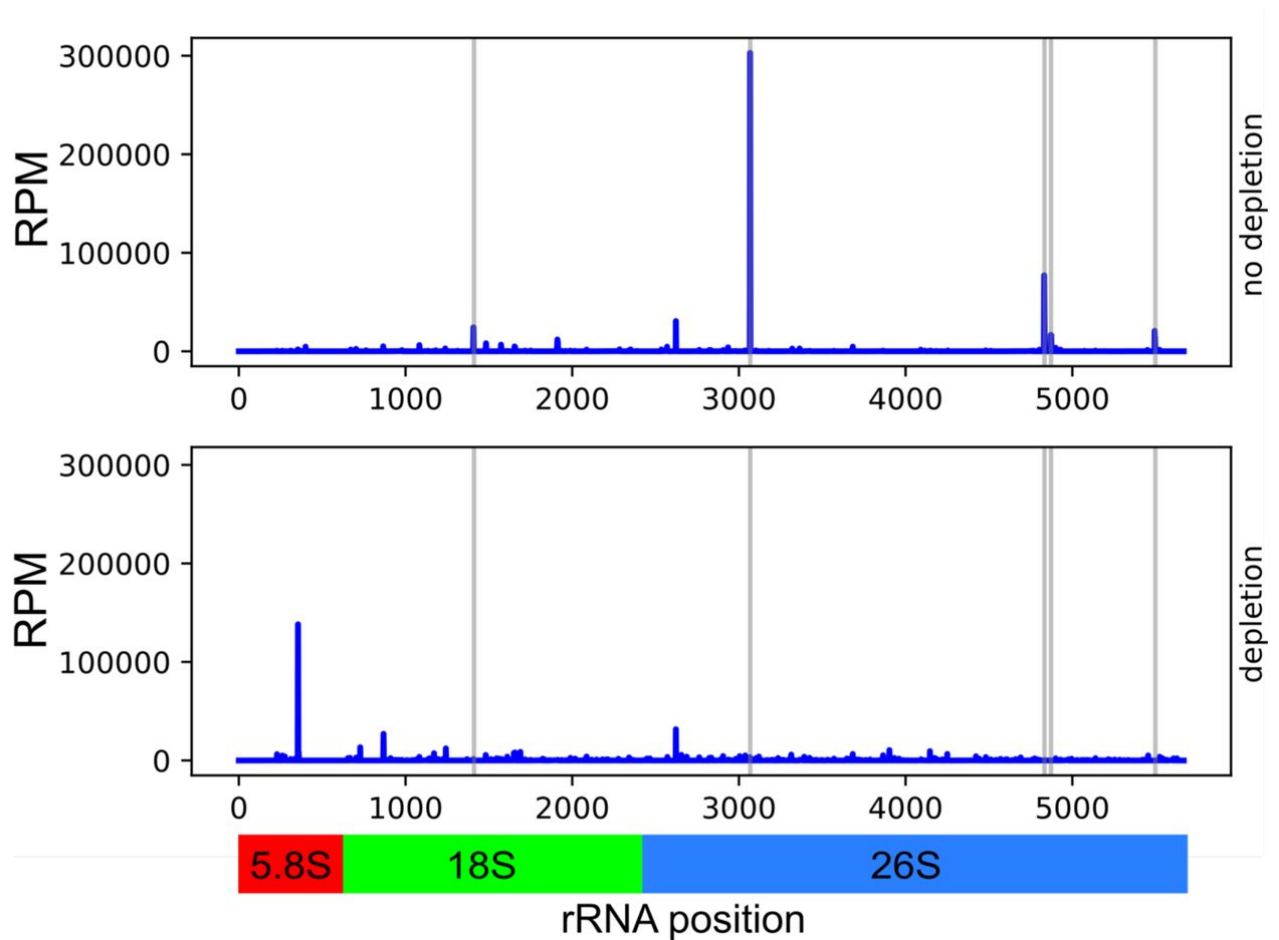
**Figure 1.** Summary of the ribosome profiling workflow. This summary is broken into five parts including lysate preparation, RPF generation and purification, library generation, data processing and data analysis. Lysate preparation includes culturing, lysis and the quantification of total RNA in a lysate. RNase digestion, monosome isolation and RPF purification represents the generation of ribosome protected fragments (RPFs). Library generation involves the conversion of small RNAs (RPFs) to a cDNA library, ready to be sequenced on an Illumina sequencing platform. Data processing involves removing of the sequencing adapters to leave only RPF sequences which are aligned to the genome. Data analysis typically involves visualizing of data via genome and/or transcriptome browser, differential gene expression and a range of others as listed in figure.



**Figure 2.** Ribosome Profiling Data from *K. marxianus*. A. Pearson's correlation of biological replicates for each experimental condition. Axis values represent log2 read counts. B. Composition of ribosome profiling library with rRNA depletion. C. Triplet periodicity of aligned RPFs for each read length. D and E display a metagene profile of aligned RPFs near the start codon and stop codon, respectively.

were obtained. First, the degree of correlation of the number of mapped reads per gene between biological replicates for each condition was assessed and found to be high with a Pearson's correlation of >0.96 (**Fig. 2A**). Second, we checked whether the RPFs actually represented known protein coding genes (**Fig. 2B**). We found that 14% of total reads aligning to the genome represented uniquely mapping RPFs; only ~0.5% of reads represented ambiguous RPFs, aligning to more than one location on the genome/transcriptome; and ~85% of the reads mapped to rRNA encoding genes. Third, we examined whether our data showed the distinctive triplet periodicity (or sub-codon phasing) of the aligned reads reflecting the 'codon-wise' movement of elongating ribosomes that is seen in ribosome profiling data. In the dataset, footprints of length 28nt (approximately half of total footprints) displayed a remarkable strong periodicity signal with ~95% of RPFs in phase with one of the three sub-codon positions (**Fig. 2C**). Finally, RPFs are expected to be massively enriched in CDS regions of genes. Using metagene profiles, we found that RPFs are largely present with CDS regions (**Fig. 2D and 2E**). In combination, these data demonstrate that the protocol generates robust ribosome profiling data.

Despite the oligo rRNA depletion, in our dataset from the large-scale experiment, ~85% of the total reads were rRNA fragments. To determine the efficiency of the targeted rRNA contamination depletion, specific rRNA contaminant sequences were analysed before and after depletion. After depletion, we see almost 100% efficiency in removal of targeted rRNA contaminants, this is visualised in **Fig. 3** where we show efficient reduction of these specific rRNA reads mapping to specific targets of rRNA. It is important to note that the introduction of rRNA contaminants can vary due to slicing of RPFs from size selection gels by free-hand, therefore rRNA abundance and composition may vary between samples and experiments. In our data, we observe this phenomenon where a sequence originating from the 5.8S is present in the post-depletion data but not in the pre-depletion data. If desired, more oligonucleotides could be designed to further reduce rRNA contamination, thus increasing the proportion of RPFs in the sequencing pool. It was interesting to note that while ambiguously mapped reads can represent >10 % of all reads in many studies from *S. cerevisiae* (seen looking at data in the Trips-Viz browser; <https://trips.ucc.ie/>) these comprised <1% of all reads in *K. marxianus*. Ambiguous mapping, whereby an RPF maps to two or more loci in the genome or transcriptome, arises because of the very short reads generated by ribosome profiling. As a result, it is not possible to determine the origin of the reads and these are generally



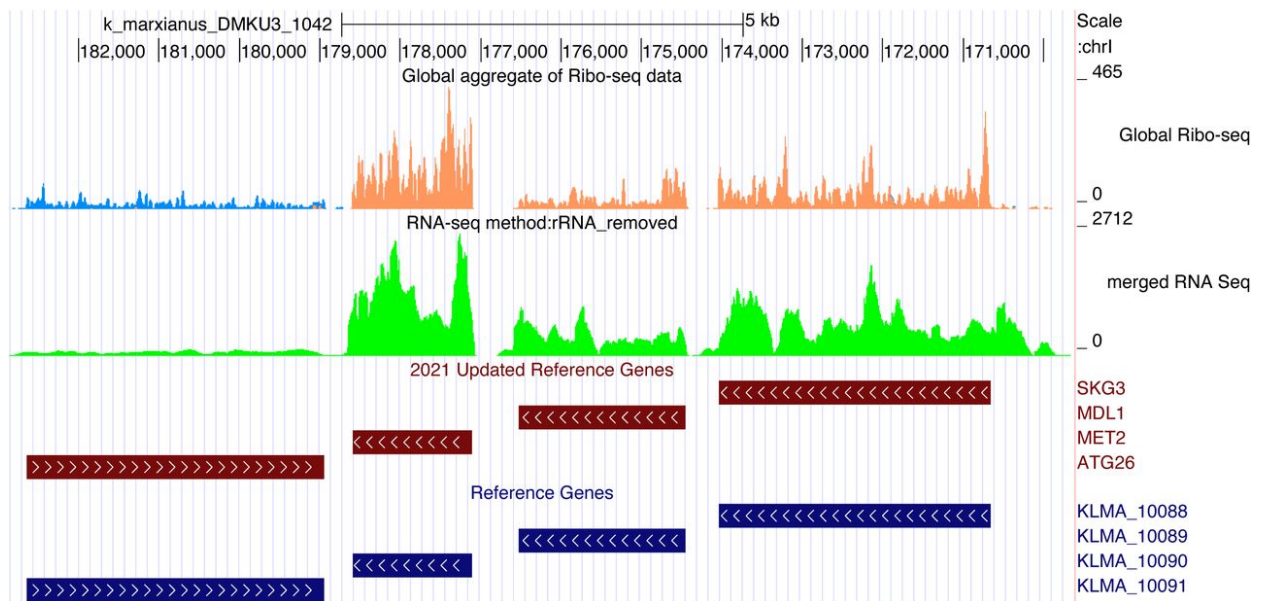
**Figure 3.** Targeted removal of nuclear encoded rRNA contaminants. Abundance is represented in reads per million (RPM) and position is relative to the generated rRNA index presented in the bottom track. Top panel represents rRNA composition and abundance with no targeted rRNA depletion employed. Bottom panel represents rRNA composition and abundance with targeted rRNA depletion protocol. Targets for rRNA depletion are highlighted as dark grey areas. Abundance is represented in reads per million (RPM) and position is relative to the generated rRNA index presented in the bottom track. The grey vertical lines highlight the rRNA contaminants that are targeted in the oligo depletion step.

discarded/ignored. This difference is most likely due to the large number of paralogous genes in *S. cerevisiae*, which arose through the proposed whole genome duplication/hybridization (WGD) event in the evolutionary history of this species (Wolfe and Shields 1997; Marcet-Houben and Gabaldón 2015). As *K. marxianus* is a pre-WGD yeast, the same issue does not apply.

### Visualisation of *K. marxianus* ribosome profiling data on public browsers

Visualisation of ribosome profiling data is important to examine translation/transcription of particular loci of interest. We previously developed two tools to allow visualisation of these data at a genome level (via GWIPS-viz) (Michel *et al.* 2014), and at the level of individual

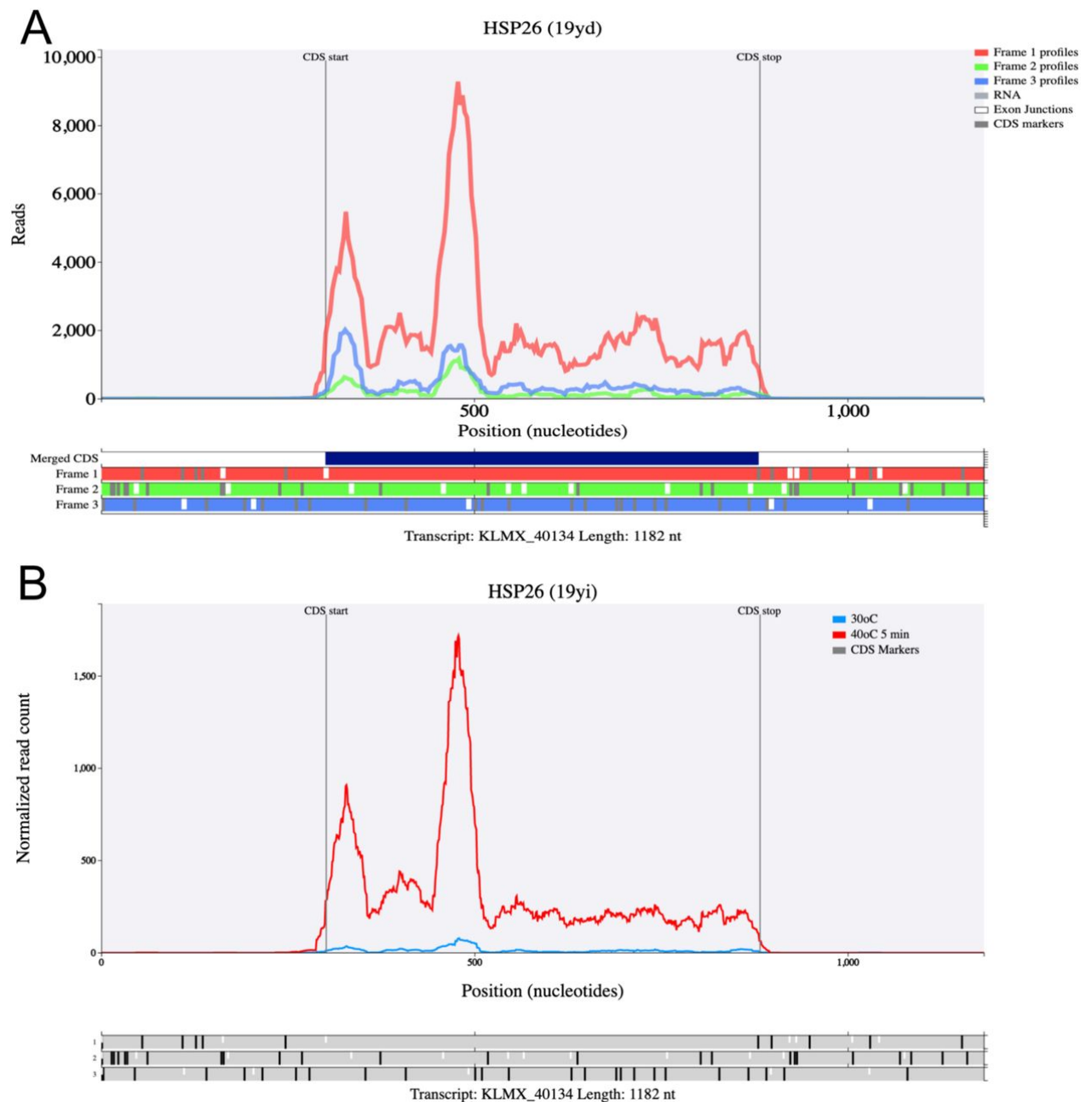
RNAs (via Trips-Viz) (Kiniry *et al.* 2019, 2021). These tools are freely accessible via RiboSeqOrg portal at <https://riboseq.org>. GWIPS-viz is a genome browser that displays RPFs mapped to each chromosome of a reference genome (Michel *et al.* 2014). The GWIPS-viz database already contained reference genomes for ~24 animal, plant, protozoal, fungal and viral genomes and we added *K. marxianus* using the genome sequence and annotation from *K. marxianus* DMKU3-1042 strain as this was the most complete genome sequence available (Lertwattanasakul *et al.* 2015). It is possible to search GWIPS-viz by gene name or gene ID and to zoom in / out of loci and as an extra feature that is new to GWIPS-viz, we included strand orientation of our ribosome profiling data to allow users determine the strand to which a RPF is mapped (orange for +/forward strand, blue for -/negative strand) (**Fig. 4**). The browser is free to use and any user that generates their own ribosome profiling data or RNA-Seq tracks (bigWigs) can upload those data as custom tracks that can be viewed privately or made public. Once uploaded, a user is able to visualise and analyse their data using all the functionality of GWIPS-viz.



**Figure 4.** GWIPS-viz Browser screenshot surrounding the *SKG3*, *MDL1*, *MET2* and *ATG26* locus of chromosome 1. Arrows on Reference Gene bars represent strand orientation. For Ribosome Profiling, orange reads represent positive strand RPFs while blue reads represent negative strand RPFs.

GWIPS-viz is mainly designed for analysis at a global level, allowing users to visualise any part of a genome, regardless of whether or not it is included in the annotations. In contrast, a second tool Trips-Viz, is a transcriptome level browser that focuses on individual mRNAs

and allows a deep analysis of translation of each mRNA (Kiniry *et al.* 2019, 2021). This transcriptome browser allows users to generate single transcript plots displaying the open reading frame that is being translated. It also allows users to visualise the distribution of RPFs along an individual mRNA while also utilising the triplet periodicity signal and differential colouring to identify potential translation in each open reading frame. As Trips-Viz did not include a reference transcriptome for *K. marxianus*, we created this reference transcriptome using our data. The application of Trips-Viz to study an individual mRNA is illustrated with an analysis of *HSP26*, using the (ribosome profiling) data for translation at 30°C and 40°C (**Fig. 5**). The top panel uses aggregate data and shows the distribution of RPFs between each reading frame and across the transcript. It is clear that reads from the first open reading frame (red) dominate, which match the position and frame of the annotated CDS. The increase in the number of reads (RPFs) at certain positions indicates ribosome stalling during translation; for example, at difficult to translate codons. The bottom panel compares the normalised read count of the correct open reading frame between the samples coming from cells grown at 30°C and 40°C. The huge increase in translation at 40°C is evident. This was to be expected as *HSP26* encodes a heat shock protein and is strongly transcriptionally induced by temperature shift. Thus, the increase in translation in this case is due to an increase in mRNA abundance. Although not shown in this simple example, in addition to single transcript plots, the Trips-Viz browser contains a large amount of metadata analyses such as triplet periodicity, read breakdowns, metaplot, protein count tables, differential expression analyses that is useful for detailed studies of translation and its regulation (Kiniry *et al.* 2019, 2021).



**Figure 5.** Trips-Viz transcript plots of *HSP26*. A. Ribosome profiling coverage for aggregated data is displayed. Note the dominant red line corresponds to the *HSP26* CDS region. B. Normalized transcript comparison plot of *HSP26* showing increased mRNA translation at 30°C and 40°C temperatures. Unique Trips-Viz plot identifiers are presented as “19yd” and “19yi” for A and B, respectively. These identifiers can be added the end of the following link to regenerate these specific plots (<https://trips.ucc.ie/short/>) on a web browser (example for panel A, <https://trips.ucc.ie/short/19yd>).

### Integrated omics studies with *K. marxianus*

Ultimately, a full suite of omics technologies, ranging for genomics to proteomics, is a requisite for comprehensive studies of any microbe. Analysis of genome sequences and transcriptomes is now relatively straightforward for diverse yeasts, but the development of



other tools still lags. Now, with the laboratory and *in silico* methods that we developed, it is possible to perform ribosome profiling with a non-model yeast, *K. marxianus*. By including both transcriptome (RNA-Seq) and translome (ribosome profiling) analysis in future studies, it will be possible to generate a comprehensive view of gene expression at a point of time and in response to a perturbation. This can be very useful to understand biological processes and for the development of strains for biotechnology. The strategy taken, and the pipeline used, can also serve as a prototype for the development of ribosome profiling methods for other yeast of biological and biotechnological interest. To facilitate the application of the tools by as many users as possible, a comprehensive step by step protocol is provided as a supplementary protocol.

### **Data Accessibility**

Ribosome profiling and RNA-Seq datasets have been deposited to the European Nucleotide Archive under the project accession number PRJEB45612. The data have also been deposited to GWIPS-viz <https://gwips.ucc.ie/> and to Trips-Viz <https://trips.ucc.ie/>.

## References

- Alva TR, Riera M, Chartron JW. Translational landscape and protein biogenesis demands of the early secretory pathway in *Komagataella phaffii*. *Microb Cell Fact* 2021;**20**:19.
- Andreev DE, O'Connor PBF, Loughran G *et al*. Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Res* 2017;**45**:513–26.
- Arevalo-Villena M, Briones-Perez A, Corbo MR *et al*. Biotechnological application of yeasts in food science: Starter cultures, probiotics and enzyme production. *J Appl Microbiol* 2017;**123**:1360–72.
- Blevins WR, Tavella T, Moro SG *et al*. Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. *Sci Rep* 2019;**9**:11005.
- Brar GA, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* 2015;**16**:651–64.
- Broach JR. Nutritional Control of Growth and Development in Yeast. *Genetics* 2012;**192**:73 LP – 105.
- Cernak P, Estrela R, Poddar S *et al*. Engineering *Kluyveromyces marxianus* as a Robust Synthetic Biology Platform Host. *MBio* 2018;**9**, DOI: 10.1128/mBio.01410-18.
- Chomczynski P, Sacchi N. The single-step method of RNA isolation by acid guanidinium thiocyanate–phenol–chloroform extraction: twenty-something years on. *Nat Protoc* 2006;**1**:581–5.
- Coloretti F, Chiavari C, Luise D *et al*. Detection and identification of yeasts in natural whey starter for Parmigiano Reggiano cheese-making. *Int Dairy J* 2017;**66**:13–7.
- Delvigne F, Zune Q, Lara AR *et al*. Metabolic variability in bioprocessing: implications of microbial phenotypic heterogeneity. *Trends Biotechnol* 2014;**32**:608–16.
- Doughty TW, Domenzain I, Millan-Oropeza A *et al*. Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts. *Nat Commun* 2020;**11**:2144.
- Duncan CDS, Mata J. The translational landscape of fission-yeast meiosis and sporulation. *Nat Struct Mol Biol* 2014;**21**:641–7.
- Eisenberg AR, Higdon AL, Hollerer I *et al*. Translation Initiation Site Profiling Reveals Widespread Synthesis of Non-AUG-Initiated Protein Isoforms in Yeast. *Cell Syst* 2020;**11**:145-160.e5.
- Fonseca GG, Heinzle E, Wittmann C *et al*. The yeast *Kluyveromyces marxianus* and its biotechnological potential. *Appl Microbiol Biotechnol* 2008;**79**:339–54.

- Fu X, Li P, Zhang L *et al.* Understanding the stress responses of *Kluyveromyces marxianus* after an arrest during high-temperature ethanol fermentation based on integration of RNA-Seq and metabolite data. *Appl Microbiol Biotechnol* 2019;**103**:2715–29.
- Gao J, Yuan W, Li Y *et al.* Transcriptional analysis of *Kluyveromyces marxianus* for ethanol production from inulin using consolidated bioprocessing technology. *Biotechnol Biofuels* 2015;**8**:115.
- Gibson UE, Heid CA, Williams PM. A novel method for real time quantitative RT-PCR. *Genome Res* 1996;**6**:995–1001.
- Groeneveld P, Stouthamer AH, Westerhoff H V. Super life--how and why “cell selection” leads to the fastest-growing eukaryote. *FEBS J* 2009;**276**:254–70.
- Hahn S, Young ET. Transcriptional Regulation in *Saccharomyces cerevisiae*: Transcription Factor Regulation and Function, Mechanisms of Initiation, and Roles of Activators and Coactivators. *Genetics* 2011;**189**:705 LP – 736.
- Hinnebusch AG. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol* 2005;**59**:407–50.
- Ingolia NT, Brar GA, Rouskin S *et al.* The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* 2012;**7**:1534–50.
- Ingolia NT, Ghaemmaghami S, Newman JRS *et al.* Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 2009;**324**:218–23.
- Ingolia NT, Hussmann JA, Weissman JS. Ribosome Profiling: Global Views of Translation. *Cold Spring Harb Perspect Biol* 2019;**11**, DOI: 10.1101/cshperspect.a032698.
- Jiménez-Gutiérrez E, Alegría-Carrasco E, Sellers-Moya Á *et al.* Not just the wall: the other ways to turn the yeast CWI pathway on. *Int Microbiol* 2020;**23**:107–19.
- Karim A, Gerliani N, Aider M. *Kluyveromyces marxianus*: An emerging yeast cell factory for applications in food and biotechnology. *Int J Food Microbiol* 2020;**333**:108818.
- Kiniry SJ, Judge CE, Michel AM *et al.* Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. *Nucleic Acids Res* 2021;**49**:W662–70.
- Kiniry SJ, O’Connor PBF, Michel AM *et al.* Trips-Viz: a transcriptome browser for exploring Ribo-Seq data. *Nucleic Acids Res* 2019;**47**:D847–52.
- Kwon D-H, Park J-B, Hong E *et al.* Ethanol production from xylose is highly increased by the *Kluyveromyces marxianus* mutant 17694-DH1. *Bioprocess Biosyst Eng*

- 2019;**42**:63–70.
- de la Torre-Ruiz MA, Pujol N, Sundaran V. Coping with oxidative stress. The yeast model. *Curr Drug Targets* 2015;**16**:2–12.
- Lane MM, Morrissey JP. *Kluyveromyces marxianus*: A yeast emerging from its sister's shadow. *Fungal Biol Rev* 2010;**24**:17–26.
- Langmead B, Trapnell C, Pop M *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25.
- Lertwattanasakul N, Kosaka T, Hosoyama A *et al*. Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. *Biotechnol Biofuels* 2015;**8**:47.
- Liu Y, Nielsen J. Recent trends in metabolic engineering of microbial chemical factories. *Curr Opin Biotechnol* 2019;**60**:188–97.
- Ljungdahl PO, Daignan-Fornier B. Regulation of Amino Acid, Nucleotide, and Phosphate Metabolism in *Saccharomyces cerevisiae*; *Genetics* 2012;**190**:885 LP – 929.
- Marcet-Houben M, Gabaldón T. Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLOS Biol* 2015;**13**:e1002220.
- Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 2011;**17**:10–2.
- Martínez-Montañés F, Pascual-Ahuir A, Proft M. Toward a genomic view of the gene expression program regulated by osmostress in yeast. *OMICS* 2010;**14**:619–27.
- Masser AE, Ciccarelli M, Andréasson C. Hsf1 on a leash – controlling the heat shock response by chaperone titration. *Exp Cell Res* 2020;**396**:112246.
- McGlinchy NJ, Ingolia NT. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* 2017;**126**:112–29.
- McManus CJ, May GE, Spealman P *et al*. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res* 2014;**24**:422–30.
- Merrick WC. eIF4F: A Retrospective \*. *J Biol Chem* 2015;**290**:24091–9.
- Michel AM, Fox G, M Kiran A *et al*. GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res* 2014;**42**:D859-64.
- Mo W, Wang M, Zhan R *et al*. *Kluyveromyces marxianus* developing ethanol tolerance during adaptive evolution with significant improvements of multiple pathways.

- Biotechnol Biofuels* 2019;**12**:63.
- Mohammad F, Green R, Buskirk AR. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *Elife* 2019;**8**, DOI: 10.7554/eLife.42591.
- Monteuuis G, Miścicka A, Świrski M *et al.* Non-canonical translation initiation in yeast generates a cryptic pool of mitochondrial proteins. *Nucleic Acids Res* 2019;**47**:5777–91.
- Morano KA, Grant CM, Moye-Rowley WS. The Response to Heat Shock and Oxidative Stress in *Saccharomyces cerevisiae*. *Genetics* 2012;**190**:1157 LP – 1195.
- de Nadal E, Posas F. Multilayered control of gene expression by stress-activated protein kinases. *EMBO J* 2010;**29**:4–13.
- Nandy SK, Srivastava RK. A review on sustainable yeast biotechnological processes and applications. *Microbiol Res* 2018;**207**:83–90.
- Parapouli M, Vasileiadis A, Afendra A-S *et al.* *Saccharomyces cerevisiae* and its industrial applications. *AIMS Microbiol* 2020;**6**:1–31.
- Rajkumar AS, Morrissey JP. Rational engineering of *Kluyveromyces marxianus* to create a chassis for the production of aromatic products. *Microb Cell Fact* 2020;**19**:207.
- Rajkumar AS, Varela JA, Juergens H *et al.* Biological Parts for *Kluyveromyces marxianus* Synthetic Biology. *Front Bioeng Biotechnol* 2019;**7**:97.
- Sanz AB, García R, Rodríguez-Peña JM *et al.* The CWI Pathway: Regulation of the Transcriptional Adaptive Response to Cell Wall Stress in Yeast. *J fungi (Basel, Switzerland)* 2017;**4**, DOI: 10.3390/jof4010001.
- Schabort DTWP, Letebele PK, Steyn L *et al.* Differential RNA-seq, Multi-Network Analysis and Metabolic Regulation Analysis of *Kluyveromyces marxianus* Reveals a Compartmentalised Response to Xylose. *PLoS One* 2016;**11**:e0156242.
- Schena M, Shalon D, Davis RW *et al.* Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;**270**:467–70.
- Sharma P, Wu J, Nilges BS *et al.* Humans and other commonly used model organisms are resistant to cycloheximide-mediated biases in ribosome profiling experiments. *Nat Commun* 2021;**12**:5094.
- Smith JE, Alvarez-Dominguez JR, Kline N *et al.* Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. *Cell Rep* 2014;**7**:1858–66.

- Spealman P, Naik AW, May GE *et al.* Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. *Genome Res* 2018;**28**:214–22.
- Steitz JA. Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. *Nature* 1969;**224**:957–64.
- Sui Y, Wisniewski M, Droby S *et al.* Responses of yeast biocontrol agents to environmental stress. *Appl Environ Microbiol* 2015;**81**:2968–75.
- Takors R. Scale-up of microbial processes: impacts, tools and open questions. *J Biotechnol* 2012;**160**:3–9.
- Taymaz-Nikerel H, Cankorur-Cetinkaya A, Kirdar B. Genome-Wide Transcriptional Response of *Saccharomyces cerevisiae* to Stress-Induced Perturbations. *Front Bioeng Biotechnol* 2016;**4**:17.
- Verduyn C, Postma E, Scheffers WA *et al.* Effect of benzoic acid on metabolic fluxes in yeasts: A continuous-culture study on the regulation of respiration and alcoholic fermentation. *Yeast* 1992;**8**:501–17.
- Wang D, Wu D, Yang X *et al.* Transcriptomic analysis of thermotolerant yeast *Kluyveromyces marxianus* in multiple inhibitors tolerance. *RSC Adv* 2018;**8**:14177–92.
- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
- Wehrs M, Tanjore D, Eng T *et al.* Engineering Robust Production Microbes for Large-Scale Cultivation. *Trends Microbiol* 2019;**27**:524–37.
- Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 1997;**387**:708–13.

## **Chapter 3 - Integrated data-driven reannotation of the *Kluyveromyces marxianus* genome reveals an expanded protein coding repertoire**

*This chapter has been deposited to BioRxiv (2022)*

<https://www.biorxiv.org/content/10.1101/2022.02.06.478964v1>

### **Abstract**

Many new biotechnology applications make use of non-traditional yeast species that possess characteristics and traits that are advantageous in specific settings. To a large extent, this is possible because of advances in molecular and genomic technologies that provide opportunities to study and manipulate the genomes of diverse yeasts. There remains, however, a substantial knowledge gap between what we know about well-studied models versus emerging industrial yeast species. In this study, we applied a multi-faceted omics analysis to better understand genome structure and gene expression in *Kluyveromyces marxianus*, a yeast widely used in food and industrial biotechnology. We combined advanced transcriptomics techniques for mapping 5' and 3' ends of RNA transcripts with ribosome profiling to explore the transcriptional and translational landscapes of this yeast. This allowed us to improve the genome annotation and identify over 300 un-annotated or mis-annotated genes. We discovered numerous examples of novel proteoforms due to use of alternative transcription or translation start sites, many genes with translated upstream open reading frames (uORFs), novel instances of frame-shifting, and other phenomena. In some cases, findings matched those for orthologous genes in *S. cerevisiae* but there were also many cases of differences, thereby creating an opportunity to explore the evolution of gene regulation in pre- and post- WGD yeasts. The processed data has been made available on the GWIPS-viz and Trips-Viz browsers, thus providing an accurate annotation of transcripts and their protein coding regions along with quantitative information on their transcription and translation.

## Introduction

*Kluyveromyces marxianus* is a budding yeast in the Saccharomycetaceae family. Although readily isolated from decaying plant matter, it is believed that some *K. marxianus* lineages were domesticated by early dairy farmers several thousand years ago and it is commonly associated with traditional fermented dairy products (Ortiz-Merino *et al.* 2018; Varela *et al.* 2019). More recently, its capacity for rapid growth, thermotolerance and other relevant traits has garnered much attention in the biotechnology sector (Fonseca *et al.* 2008; Lane and Morrissey 2010; Morrissey *et al.* 2015; Varela *et al.* 2017; Karim, Gerliani and Aider 2020). This has led to substantial progress in the development of gene engineering and synthetic biology tools (Nambu-Nishida *et al.* 2017; Cernak *et al.* 2018; Rajkumar *et al.* 2019; Rajkumar and Morrissey 2022), meaning that it is now relatively straightforward to reprogramme strains as cell factories for industrial biotechnology applications (Liu and Nielsen 2019; Rajkumar and Morrissey 2020; Baptista, Cunha and Domingues 2021; Patra *et al.* 2021). Other developments at the physiological level are improving the potential for large-scale fermentation of *K. marxianus* under industrial conditions (Dekker *et al.* 2021). There have also been several genome-wide studies that explored gene expression at the transcriptional level (reviewed in (Ha-Tran, Nguyen and Huang 2020).

To date, 17 *K. marxianus* genomes have been published, with 3 strains (NBRC 1777, DMKU-1042 and FIM1) having fully sequenced genomes, annotated and assembled on a chromosomal level (Inokuma *et al.* 2015; Lertwattanasakul *et al.* 2015; Mo *et al.* 2019). Similar to other pre-whole genome duplication yeast in the Saccharomycetaceae, *K. marxianus* possesses eight nuclear chromosomes and a haploid genome of ~10.9 Mb in size, although variations in ploidy and aneuploidy are common in the domesticated dairy strains (Fasoli *et al.* 2015; Ortiz-Merino *et al.* 2018). Annotation of the sequenced genomes indicates that there are ~5000 protein-coding genes in *K. marxianus*, lower than the ~6000 in *S. cerevisiae*. There have been only limited efforts to explore the entire *K. marxianus* genome to establish the basis of its unique or interesting traits. Most previous studies were narrowly focused, for example on the expansion and diversification of sugar transporters (Knoshaug *et al.* 2015; Varela *et al.* 2017, 2019; Donzella *et al.* 2021). In a recent comparative omics study, we discovered that genes that were evolutionarily young and / or unique to *K. marxianus* are overrepresented among genes that display differential expression when growing under stressful conditions (Doughty *et al.* 2020). We also established that at least



one of these genes is required for competitive growth at high temperature (Montini *et al.* 2022). These findings reinforce the need to thoroughly explore and characterise the *K. marxianus* genome as many of the unique phenotypic traits in *K. marxianus* could be due to genes not yet characterised in any species, or indeed to known genes that are regulated differently or encode proteins with alternative functionality.

Genome annotation in yeast is generally homology-based, which is rapid but brings some limitations, especially when dealing with less well-characterised species. Among the limitations of this conventional protein-coding gene annotation approach, is the difficulty in considering the diversity of proteoforms that can be encoded within a single gene locus. This is illustrated well in *Saccharomyces cerevisiae* where studies revealed a range of different types of proteoforms generated with alternative translation initiation codons (Monteuuis *et al.* 2019), translational readthrough (Namy *et al.* 2003), and frameshifting (Atkins *et al.* 2016). Proteoforms with different N-termini can arise through either transcription or translation-based mechanisms and, amongst other roles, are known to have different localisation within the cell, due to a targeting signal present at the N-terminus of the longer proteoform. For example, *S. cerevisiae* *HTS1* uses alternative transcription start sites that give rise to long and short mRNA isoforms that are then translated from different start codons 60 nt apart (Natsoulis, Hilger and Fink 1986). In contrast, translation of the *S. cerevisiae* *ALA1* mRNA involves leaky scanning, whereby a proportion of ribosomes initiate translation at upstream near-cognate start codons (ACG), while remaining ribosomes continue scanning and initiate translation downstream at the main AUG start codon (Tang *et al.* 2004). Other translational mechanisms can result in proteoforms with alternative C-termini. This includes stop codon readthrough, whereby ribosomes fail to terminate at stop codons and continue translation, or ribosome frameshifting, whereby a ribosome repositions and translates in the -1 or +1 open reading frame (ORF) (Namy *et al.* 2003; Atkins *et al.* 2016). In addition to enabling synthesis of alternative proteoforms, these mechanisms could provide regulatory sensing of cellular conditions. The best-known example is the regulation of intracellular levels of polyamines via +1 frameshifting that takes place during translation of *OAZ1* mRNA (Palanimurugan *et al.* 2004; Ivanov, Gesteland and Atkins 2006), which is conserved from yeast to humans (Ivanov and Atkins 2007). Other important translated regions that can be missed in gene annotation include small upstream open reading frames (uORFs) within 5' leaders of mRNAs. In *S. cerevisiae* these uORFs have been shown to regulate translation of mRNAs

including for example *CPA1*, which encodes an arginine attenuator peptide, where ribosome stalling at an uORF is regulated via intracellular arginine levels (Gaba *et al.* 2001; Gaba, Jacobson and Sachs 2005).

Regarding studies of gene expression, massively parallel sequencing facilitated a switch in research focus from gene-specific studies to the genome-wide scale. Transcriptomic methods such as RNA-seq can be used to measure gene expression (Wang, Gerstein and Snyder 2009) or to capture transcript start and polyadenylation sites, revealing alternative mRNA isoforms based on 5' or 3' ends (Adiconis *et al.* 2018; Yu *et al.* 2020). Ribosome profiling is another genome-wide tool to measure gene expression. This tool captures and locates the positions of translating ribosomes, revealing precise regions of the genome that are translated at a moment in time. Using ribosome profiling (Ribo-Seq), it is possible to investigate the translational control of specific mRNAs, measure gene expression changes at a translational level, determine the translation efficiency of a transcript, and identify novel coding regions (Ingolia *et al.* 2009; Michel and Baranov 2013; McManus *et al.* 2014; Brar and Weissman 2015; Kiniry, Michel and Baranov 2020). Ribosome profiling was first carried out in *S. cerevisiae* (Ingolia *et al.* 2009), then later in the other yeasts, *Saccharomyces paradoxus* (McManus *et al.* 2014), *Schizosaccharomyces pombe* (Duncan and Mata 2014), *Saccharomyces uvarum* (Spealman *et al.* 2018), *Komagataella phaffii* (Alva, Riera and Chartron 2021) and *Candida albicans* (Sharma *et al.* 2021). Using ribosome profiling, it has been possible to uncover many non-canonical aspects of gene structure and regulation. In *S. cerevisiae*, the approach revealed a wide-range of translated upstream open reading frames (uORFs) within 5' leaders of mRNAs (Ingolia *et al.* 2009), widespread non-AUG initiation encoding unannotated proteoforms (Monteuuis *et al.* 2019; Eisenberg *et al.* 2020) and identified a number of previously unknown translated small ORFs (Smith *et al.* 2014). In *K. marxianus*, genome-wide studies of gene expression have focused mainly on RNA-seq (reviewed in (Ha-Tran, Nguyen and Huang 2020)), ignoring these more complex mechanisms as they only reveal relative mRNA abundances.

We applied the ribosome profiling method to *K. marxianus* and established a protocol and pipeline for ribosome profiling in this yeast (Fenton *et al.*, 2022). To gain a better understanding of the *Kluyveromyces* transcriptome and translome, we employed a combination of transcriptomics techniques and ribosome profiling that allowed us to improve

its genome annotation by increasing its accuracy, and inclusion of alternative proteoforms. In addition to depositing our primary data and annotations into standard depositories we generated a *K. marxianus* entry at GWIPS-viz (Michel *et al.* 2014, 2018) and Trips-Viz (Kiniry *et al.* 2019, 2021) browsers where processed data can be freely and conveniently explored alongside the new annotation.

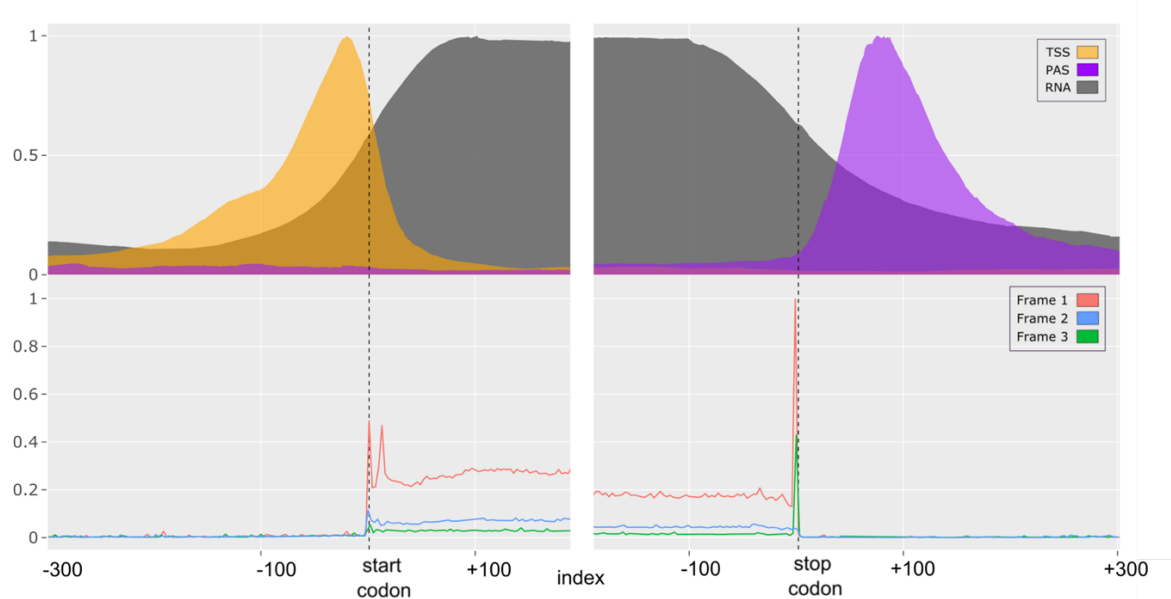
## Results

### Generation of Multiomic Data

We wanted to generate a holistic view of gene expression in *K. marxianus* to understand the coding potential of the genome and to explore its potential for variation. To do this, we employed an integrated combination of transcriptomics and ribosome profiling in what can be called a “multi-omics” analysis since a range of different techniques are used to analyse the resulting data (Figure 1). Transcriptome analysis was performed with now-standard RNA-Seq methods and for ribosome profiling in *K. marxianus*, we applied a custom protocol that is a modified version of that used in *S. cerevisiae* (McGlinchey and Ingolia 2017; Fenton *et al.* 2022). To try to capture expression of as many genes as possible, our experimental design involved the collection of ribosome profiling and RNA-seq data from cultures grown at 30°C, at 40°C. and at 5, 15, 30 and 60 minutes after a transfer from 30°C to 40°C. Using the updated gene annotation that is described later, we detected the translation of 4872 protein-coding genes with at least 20 mapped ribosome footprints (RPFs), representing ~95% of the previously annotated protein-coding genes in *K. marxianus*. Ribosome profiling data displayed excellent triplet periodicity, allowing us to interpret which frame was translated in a particular locus (Figure 1).

While ribosome profiling identified the regions of genes that were translated, we also wanted to map the 5' and 3' ends of mRNAs. Since the RNA-seq reads generated in our experiments are not suitable for accurate mapping of transcription start sites (TSS), we made use of TSS-seq data that was available from a previous study with the strain DMKU3-1042 (Lertwattanasakul *et al.* 2015). From the published data, it was possible to precisely map the transcription start sites (TSS) for 2901 *K. marxianus* genes (57% of the total). The majority of genes displayed a single TSS but 489 genes had two TSS and ~150 had three or more TSS (Supplementary Figure 1). While this level of heterogeneity is comparable with that seen in *S. cerevisiae* (Arribere and Gilbert 2013), it was found that the median 5' leader length of 97 nt

was almost twice that reported in *S. cerevisiae* (Nagalakshmi *et al.* 2008). In some cases (407 genes), TSS were detected within the coding sequence indicating that transcriptional variation can give rise to alternative shorter mRNA isoforms. To locate polyadenylation sites (PAS) on the 3' ends of mRNAs, we designed a computational approach to identify and separately align RNA-seq reads with polyA tags (using our polyA-enriched RNA-seq data, which was suitably 3' biased for this purpose). We mapped polyadenylation sites (PAS) for 4682 genes (91% of the total), of which ~3000 genes contain a single PAS and ~1700 genes had more than one PAS (Supplementary Figure 1). The median distance between the stop codon and the PAS was 128 nt, similar to values reported in *S. cerevisiae* of 104 nt (Nagalakshmi *et al.* 2008) or 166 nt (Ozsolak *et al.* 2010), indicating that the 3' trailer lengths are comparable in both species. As with TSS, PAS mapped within the CDS for some genes. Indeed, 607 genes have an internal PAS, and for 335 of these genes it is a major site for polyadenylation. The presence of 5' truncated transcript isoforms have been reported previously (Arribere and Gilbert 2013), but the exact roles of these transcripts remain elusive and the same study discovered many of these truncated mRNAs are subject to nonsense mediated decay (NMD) due to out of frame initiation and termination at premature stop codons. Together, these data revealed the genome-wide locations of transcript start sites and polyadenylation sites (Supplementary tables 1 & 2), allowing us to accurately determine the boundaries of expressed mRNAs in *K. marxianus*.



**Figure 1.** Multiomics metagene profile plot for all annotated protein coding genes relative to start and stop codons. The top plot (transcriptomic data) shows densities of reads from TSS (transcript start sites), PAS (polyadenylation sites) and RNA-seq experiments. The bottom plot shows densities of ribosome footprints differentially coloured based on the best supported reading frame with Frame 1 corresponding to the frame of annotated CDS. Vertical dashed lines represent the start and stop codons of CDS regions. Outward CDS boundaries include 300 nt from the start or stop codon. Inward CDS boundaries extend 200 nt downstream of CDS start and 200 nt upstream of CDS stop. The second red peak downstream of the start codon in the ribosome profiling track represents circularization bias (or circLigase II) due to the preferred selection of 5' A nts representing which would represent the RPFs beginning with the start codon (AUG).

### Multiomics analysis reveals potential gene regulation and expression of alternative proteoforms

Combining ribosome profiling data with accurate transcript start and polyadenylation site positions, we investigated the extent to which variability in transcription, polyadenylation and translation are responsible for the synthesis of alternative proteoforms in *K. marxianus*. We created a computational pipeline that uses ribosome profiling data to call all potential translated ORFs outside of canonical annotated protein coding genes, and then classified them depending on their position relative to known protein coding genes. Parameters such as P-site scores, ORF length, distance to parent gene and number of RPFs were used to filter the data (see Methods and Supplementary Figure 4). This pipeline identifies features such as N-terminal extensions (NTE), internal ORFs (iORF), upstream open reading frames (uORF), overlapping upstream open reading frames (ouORF), antisense translation (aORF) and ribosome frameshifting (see Supplementary Figure 5 for visualization of these ORFs). These were systematically explored on a genome-wide level and in each candidate case, manual

multiomics visualization around each locus of interest confirmed the presence of the reported feature. Full lists of genes displaying these features are provided in Supplementary tables 3 – 7 and specific examples of each are described below.

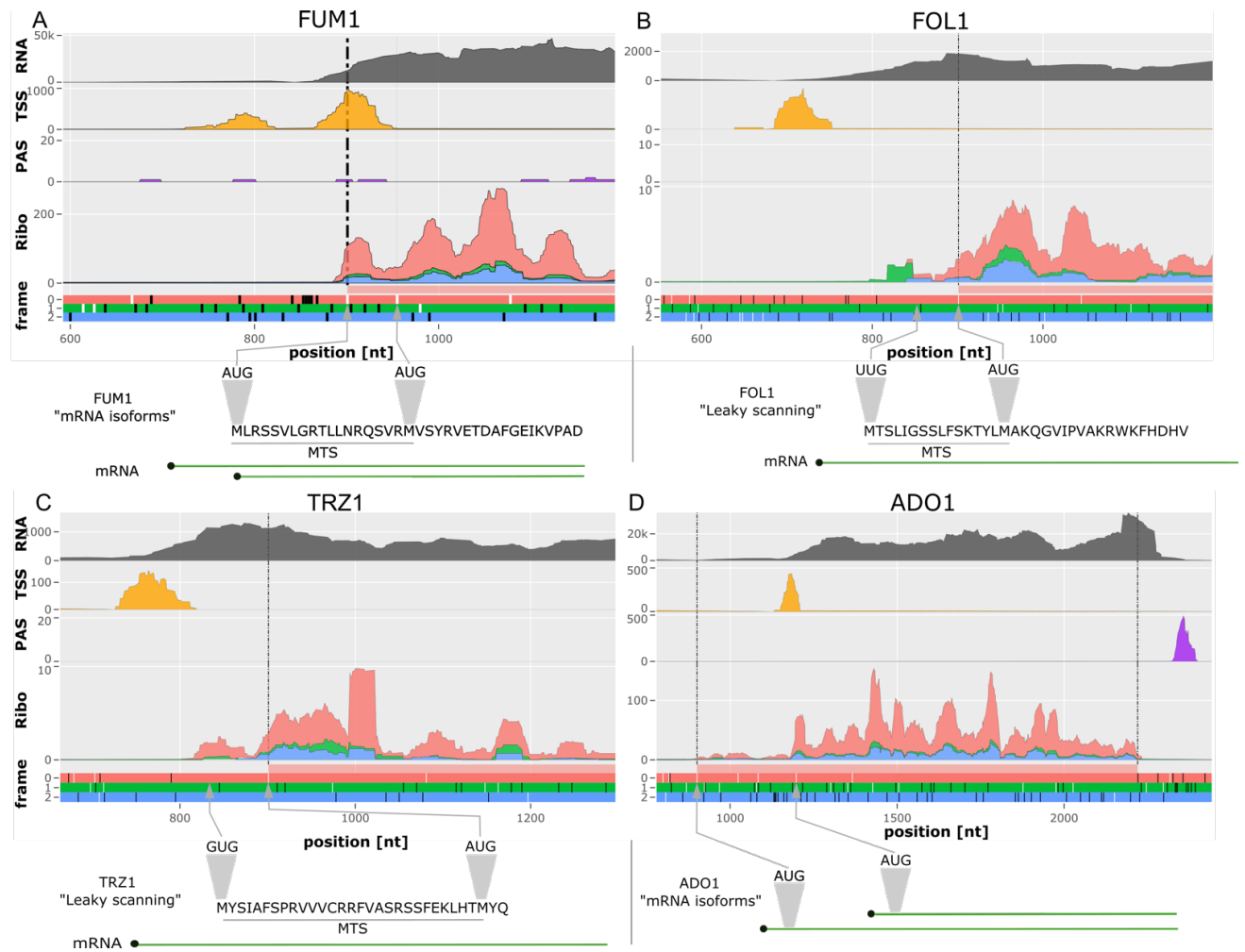
### 1. Expression of alternative proteoforms

Proteins with alternative N-termini (PANTs) can arise due to regulation at the level of transcription or translation. Use of an alternative transcription start site (aTSS) can give rise to a longer or a shorter mRNA isoform and the consequential use of a different translation start codon. In contrast, the use of alternative translation initiation sites (aTIS) arises because of recognition of different start codons, for example *via* leaky scanning. Both types of PANTs were evident in *K. marxianus* and we detected a total 5499 potential PANTs in 4825 genes (Supplementary table 3). It is known that N-terminally extended (NTE) proteoforms are sometimes used in *S. cerevisiae* to encode mitochondrial targeting signals (MTS) that localise to the mitochondrion (Monteuuis *et al.* 2019), therefore we used mitochondrial signal prediction methods to assess our NTE candidates (see methods). This analysis predicted that 322 NTEs may encode a MTS (Supplementary table 4).

Several examples of *K. marxianus* genes that use either aTSS or aTIS to generate PANTs are depicted using multi-omics plots in Figure 2. In the case of three genes (*FUM1*, *FOL1* and *TRZ1*) the PANTs control mitochondrial localisation, and for the fourth (*ADO1*), the function of the alternative isoforms is not known. *FUM1*, encodes fumarase, which can be located in either the cytoplasm or mitochondrion in *S. cerevisiae* (Wu and Tzagoloff 1987). In *K. marxianus*, two distinct TSS are visible (orange peaks), one of which overlaps the annotated start codon (Figure 2A). There is clear evidence of translation from the annotated start codon, however this is most likely to occur from the longer mRNA isoform as the position of the downstream TSS precludes use of this AUG codon for initiating ribosomes assembling at the 5' end of mRNA. There is a second in-frame AUG codon 51 nts downstream of the first, and the large increase in RPFs downstream of this AUG codon is a strong indication that translation initiates here. While it is possible that this second AUG codon is accessed *via* leaky scanning, it is not likely because the first AUG codon is in a strong Kozak context, and thus, the shorter proteoform is probably translated from an mRNA isoform made using the downstream aTSS. The peptide sequence between these two codons is predicted to be a mitochondrial targeting signal (MTS) and the data suggest that the localisation of Fum1p to

the mitochondrion or the cytoplasm is regulated at the level of transcription. This differs from *S. cerevisiae*, where it has been proposed that regulation of protein folding *via* intracellular metabolites determines localisation of Fum1 (Herrmann 2009; Regev-Rudzki *et al.* 2009). In contrast to Fum1, Fol1, encoding a multifunctional enzyme involved in folic acid biosynthesis, is an example of a protein where PANTs with, or without, an MTS arise due to the use of alternative translation start sites (aTIS) on a single transcript (Figure 2B). In this case, there is a single TSS but clear evidence of a low amount of translation initiation upstream of the annotated AUG. This translation initiates from a near-cognate UUG codon, which is likely to be inefficiently recognised, causing leaky scanning and subsequent initiation at the downstream AUG. Further confirmatory analysis of the use of these two translation start sites is presented in Supplementary Figure 6. Interestingly, Fol1 was reported to be exclusively located in the mitochondrion in *S. cerevisiae* (Guldener *et al.*, 2004), again raising questions as to possible differences between *K. marxianus* and *S. cerevisiae*. *TRZ1*, which encodes a tRNA endonuclease that is localized to both the nucleus and mitochondrion in *S. cerevisiae* (Chen *et al.* 2005; Skowronek *et al.* 2014), is another example of where PANTs arising from leaky scanning of a non-cognate start codon (Figure 2C). Again, there is a single TSS but in this case, translation starts at a GUG codon 27 codons upstream of the annotated AUG start codon, which allows incorporation of the MTS. It is noteworthy that in *S. cerevisiae*, ribosomes initiate translation at an upstream CUG codon in the *TRZ1* mRNA (Monteuuis *et al.* 2019). While the localization signals are generally short, we also find examples of PANTs with considerable length variation at their N-termini. An example is *ADO1*, encoding an adenosine kinase, where we found evidence for aTSS giving rise to PANTs differing by 100 AA (Figure 2D). Here, both RNA-seq and Ribo-seq coverage suggests the presence of two translated mRNA isoforms with the shorter isoform being more abundant. Indeed, in the figure, both the first TSS and the translation from an upstream AUG are just visible on this scale. Triplet periodicity supports the premise that the upstream AUG is used. No significant matches were found for the extended region in Conserved Domain Database (CDD), nor among protein families in InterProScan but the transmembrane topology and signal peptide predictor Phobius (Käll, Krogh and Sonnhammer 2004) detected a signal peptide present at the N-terminus of the longer proteoform Supplementary Figure 7) suggesting that the two proteoforms may have different compartmentalisation. *S. cerevisiae* Ado1 lacks this extended region, but it is present in other *Kluyveromyces* species, indicating

that the extended Ado1 variant may have a specific, though as-yet unknown, function in *Kluyveromyces* spp.



**Figure 2.** Multiomics evidence of PANTs. Upper panels represent densities of sequencing reads obtained with different techniques, coverage is displayed with the stacked method. For ribo-seq data, RPFs are differentially coloured to match reading frames in the ORF architecture plot below with dotted line showing the positions of the start of annotated CDS (in red). In ORF plots white lines represent AUG codons while black lines represent one of three stop codons. The bottom schematic illustrates suggested models of PANTs expression. A. *FUM1* exemplifies PANTs expression from two RNA isoforms where the location of the most 5' AUG codons differ. B. *FOL1* exemplifies PANTs expression from the same mRNA where initiation at two different starts (UUG and AUG) occurs due to leaky scanning. A more granulate view of translation initiation from the upstream UUG is shown in supplementary figure 6. C. *TRZ1* exemplifies PANTs expression from GUG and AUG start codons. D. *ADO1* exemplifies PANTs expression via two mRNA isoforms both utilizing AUG start codons. *ADO1* exemplifies PANTs expression via two mRNA isoforms both utilizing AUG start codons.

Translation of internal ORFs (iORFs), which arise from alternative translation of the +1 or -1 frame (relative to the AUG codon) within the annotated CDS of a gene, is another mechanism that cells use to generate alternative proteoforms. We found 861 potential instances of this in

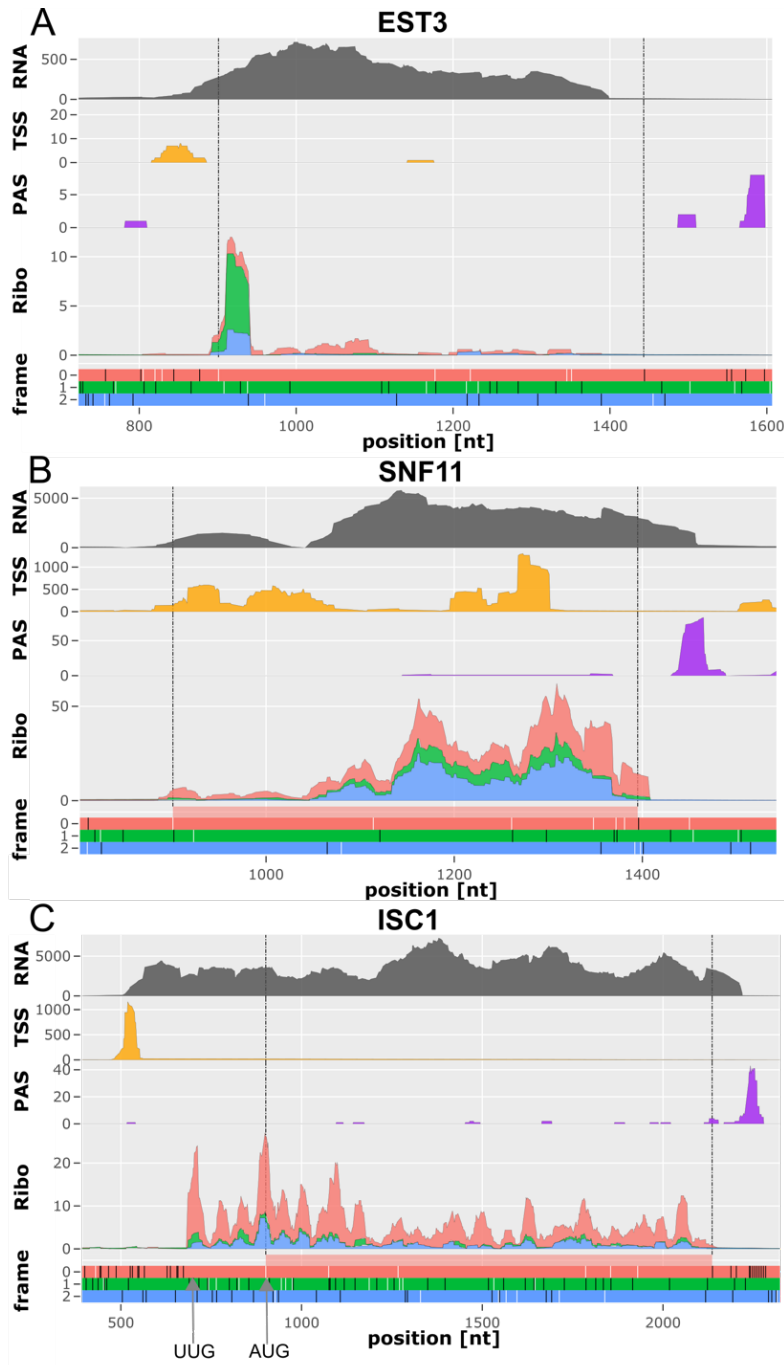


*K. marxianus* (Supplementary table 5) and examined several in more detail to illustrate the depth of data that can be retrieved from the multi-omics analysis. *EST3* is an example where such internal translation could be due to leaky scanning past the annotated start codon with initiation at a downstream AUG codon in a different frame (Figure 3A). Initiation at the downstream AUG would give rise to a 24 AA peptide whereas initiation at the first annotated AUG leads to a predicted protein that is homologous to Est3 in other yeasts. The first AUG codon is in poor context (UCCAAUGCCC), hence it is likely that a proportion of scanning ribosomes fail to recognize the main start codon and initiate at this downstream AUG. An alternative process whereby scanning ribosomes can slide from one AUG to another during the initiation process while awaiting a critical GTP hydrolysis step has been described in mammalian systems that when two AUG start codons are in close proximity, and such a mechanism cannot formally be ruled out here (Terenin *et al.* 2016). Regardless of the precise mechanism, it is seen that there is strong translation of the first ORF (Figure 3A, green RPFs) and weaker translation of the second ORF (Figure 3A, red RPFs), which encodes functional Est3. Interestingly, in *S. cerevisiae* and most other yeasts, *EST3* utilizes +1 frameshifting and it was previously noted that the *Kluyveromyces* genus is an exception that does not utilize frameshifting (Farabaugh *et al.* 2006). Both +1 frameshifting in *S. cerevisiae* and suboptimal initiation in *K. marxianus* are expected to result in low translation efficiency of the *EST3* mRNA, and it appears that different yeast lineages have arrived at distinct mechanisms to translationally restrict the level of Est3. One can speculate that this may be a regulated process whereby some stimulus would act to overcome the translational controls allowing production of higher amounts of the protein.

A different type of iORF is seen at the *SNF11* locus, which encodes a subunit of the SWI/SNF chromatin remodelling complex (Figure 3B). In this case, translation from the first (annotated) AUG gives rise to Snf11 (165 AA) but there is also evidence of both additional transcription starts and translation initiation from other AUG codons downstream of the annotated AUG. In fact, the TSS data suggest several sites of transcription initiation that would lead to the production of multiple mRNA isoforms shorter than the annotated one. These isoforms lack the annotated *SNF1* start codon and are likely to be translated from AUG codons further downstream. This idea is supported by the increased ribosome profiling density in the area of the second in-frame AUG (Figure 3A, red RPFs) as well as around a -1 frame AUG codon located a few nucleotides upstream of the second in-frame start codon

(Figure 3A, blue RPFs). Use of these alternative AUGs would give rise to N-terminally truncated proteoforms in the case of in-frame codons, and a 92 AA peptide if the -1 frame was used. For the latter, the predicted protein does not have any homologs in databases so its significance remains to be established.

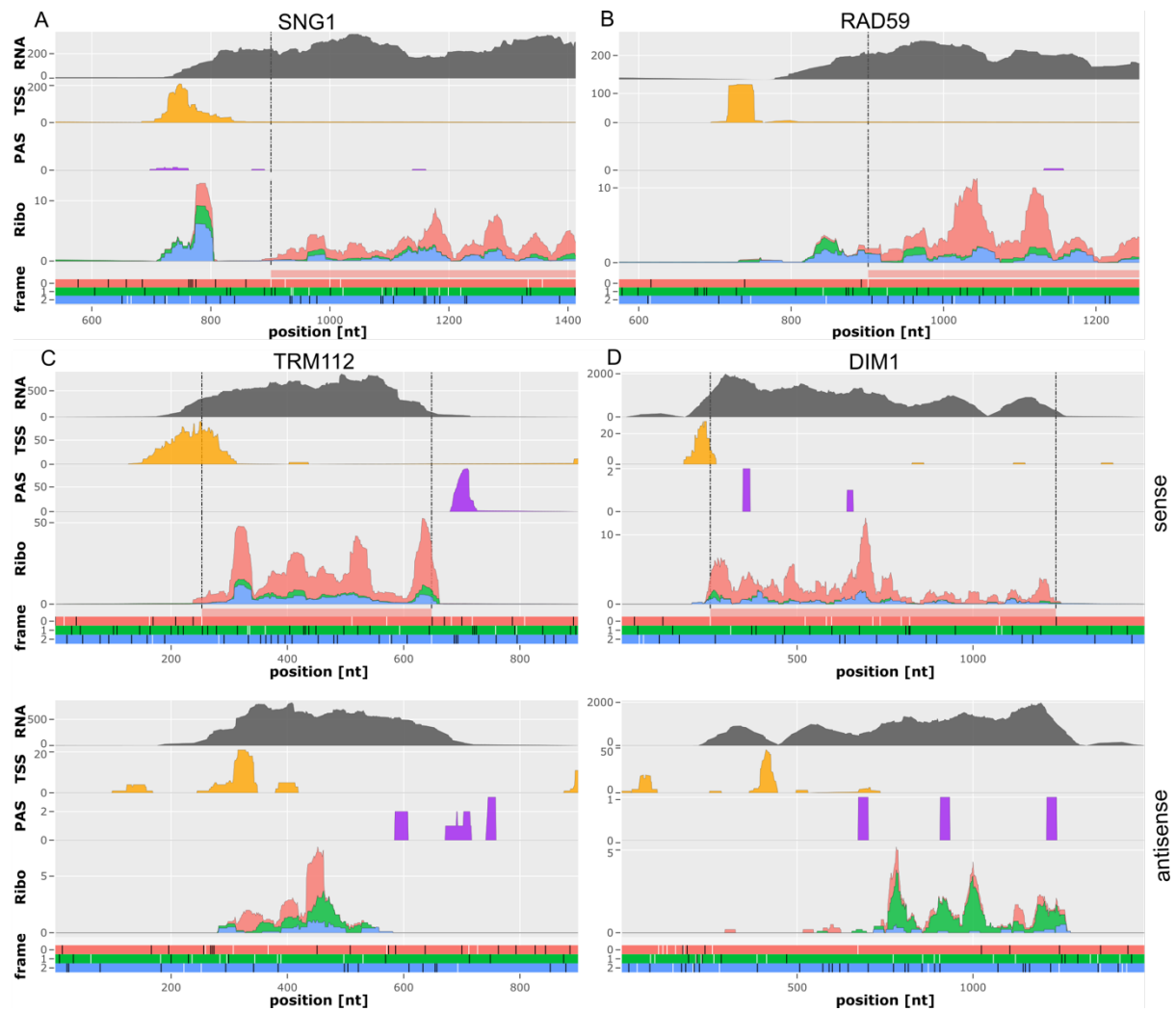
*ISC1*, encoding inositol phosphosphingolipid phospholipase C, is an intriguing case that reveals unusual use of non-AUG initiation codons in *Kluyveromyces* spp. Initial analysis suggested that this locus might encode a PANT as translation upstream of the annotated AUG start codon was evident (Figure 3C). Deeper analysis, however, revealed that translation occurs exclusively from an upstream UUG start codon, which is in good Kozak context (AAG\_UUG\_ACG). TSS and RNA-seq confirms that the *ISC1* locus encodes a single major transcript isoform. The possibility that leaky scanning could allow the annotated AUG codon be used is discounted as there are multiple of out-of-frame AUG codons between the TSS and this AUG. We excluded the possibility that the observed initiation at UUG was an artefact due to a recent mutation of the UUG to an AUG in the strain that we used to generate ribosome profiling by examining the sequence of RPFs aligned to this region and confirming that there were no mismatches. To determine if this exclusive non-AUG translation is conserved, we aligned the *ISC1* protein sequences from *K. marxianus*, *Kluyveromyces lactis* and *S. cerevisiae* and also examined the locus architecture (Supplementary Figure 8). The architecture in *K. lactis* matches *K. marxianus* but, in *K. lactis*, the predicted start codon is a GUG. The start codon Kozak context of both GUG (*K. lactis*) and UUG (*K. marxianus*) are identical (AAA\_NUG\_ACG, where N is G or U) and the N-termini of the proteins are conserved, providing a strong indication that this region is translated in both yeasts. We also used Trips-Viz and ribosome profiling data from multiple published studies at the *S. cerevisiae* *ISC1* locus to examine translation in this species (Supplementary Figure 8). There is no evidence of the use of non-AUG translation but we did observe that the start codon is incorrectly annotated in *S. cerevisiae*, with in-frame translation starting at an AUG codon downstream of the annotated start codon (see Supplementary Figure 8). For the alignments shown in Supplementary Figure 8, the previously predicted upstream 29 amino acids of the *S. cerevisiae* Isc1 were removed as these are not translated. Despite the difference in the choice of start codon, there is evidence of conservation of the N-termini between *Kluyveromyces* and *Saccharomyces*, raising further questions as to why *Kluyveromyces* has evolved to use non-cognate start codons for this gene (Supplementary Figure 9).



**Figure 3.** Multiomics plots display out of frame internal translation initiation of known genes (iORFs) and exclusive non-AUG translation initiation. A. *EST3* main CDS is encoded in red frame and described iORF is in green (+1 frame). B. *SNF11* locus is encoded in zero (red) frame while iORF is encoded in the -1 (blue) frame. C. Multiomics plot of the *ISC1* locus. The annotated CDS is in the red frame. The proposed exclusive UUG and annotated AUG start codon are marked below the frame track. See Figure 2 for explanation of the multiomics plots.

## 2. Translation of short ORFs within 5' leaders

Upstream open reading frames (uORFs) are one of the best-studied mechanisms by which translation of an mRNA is regulated in yeast and fungi (Hood *et al.* 2009). In addition, across eukaryotes, many uORFs are known to regulate translation in response to various stimuli, such as polyamines in mammals (Law *et al.* 2001; Ivanov *et al.* 2018; Vindu *et al.* 2021) and plants (Franceschetti *et al.* 2001), magnesium levels in mammals (Hardy *et al.* 2019), boron in plants (Tanaka *et al.* 2016) and arginine levels in yeast (Gaba, Jacobson and Sachs 2005) among many others. In our analysis of the *K. marxianus* data, we detected 818 uORFs that do not overlap with CDS (see Supplementary table 6). These included *GCN4*, which is considered a paradigm for translational regulation via uORFS in yeast (Hinnebusch 2005) where several short uORFS are translated, supporting the idea that the *GCN4* regulatory system in *K. marxianus* is identical to that of *S. cerevisiae* (Supplementary Figure 10). A further example to illustrate this type of uORF is presented in Figure 4A, where strong initiation at an AUG-initiated uORF ~140 nt upstream of the main CDS for *SNG1* is seen. *SNG1* encodes a protein involved in drug resistance so it also fits the pattern of this type of regulation being used for some stress-response genes (García-López *et al.*, 2010). We also identified 443 potential overlapping uORFs (ouORFs), which we define as cases where the uORF overlaps the main ORF/CDS and thus translation is expected to be mutually exclusive (Supplementary table 7). *RAD59*, encoding a DNA repair protein provides an example of this, where an ouORF (Figure 4B, blue RPFs) overlaps the main ORF (Figure 4B, red RPFs).



**Figure 4.** Translation within 5' leaders and antisense translation of protein-coding genes. A. uORF in the 5' leader of the *SNG1* mRNA in blue frame (left). B. uORF in the *RAD59* mRNA in the blue frame which overlaps the main CDS (right). Multiomics plots displaying sense multiomics of C. *TRM112* and D. *DIM1* (top row). Lower row displays exact antisense locus (5' to 3') For *TRM112*, antisense translation can be seen predominantly in two ORFs (red and green) while for *DIM1*, antisense translation mostly occurs in the green frame. See Figure 2 for explanation of the multi-omics plots.

### 3. Translation of Antisense mRNA

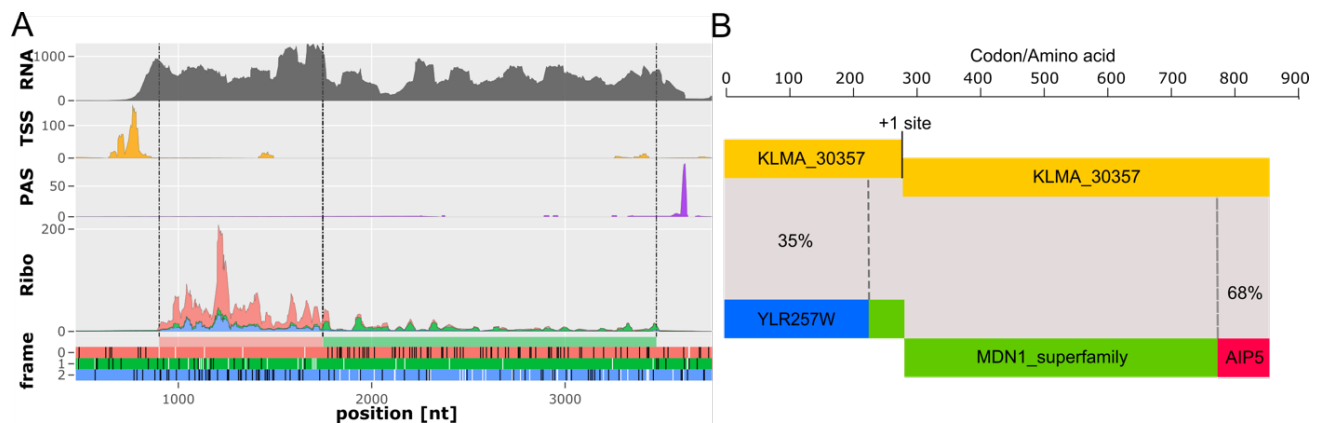
Antisense translation represents a class of detected translated ORFs (aORFs) on the opposite strand of an annotated protein coding gene. Using GWIPS-viz, we unexpectedly observed that RPFs map to sense and antisense strands of *TRM112* and *DIM1*, suggesting the existence of translated antisense transcripts (Figure 4C and 4D & Supplementary Figure 11). We therefore decided to include aORFs in our computational pipeline to find more candidates for antisense translation. We ranked genes by the number of RPFs aligned to the opposite strand of all protein-coding genes and then validated candidates using multiomics

visualisation. This approach uncovered potential antisense expression in 938 genes including *TRS23*, *VPS9*, *bioA* and *EXO84*, which were confirmed with multiomics plots and GWIPS-viz (Supplementary table 8). Antisense translation was previously noted by Duncan and Mata in ribosome profiling data from *S. pombe* (Duncan and Mata 2014) but its biological significance remains unknown.

#### 4. Novel +1 ribosomal frameshifting at the *KLMX\_30357* locus

We uncovered a potential +1 frameshifting event during translation of an mRNA derived from the *K. marxianus* *KLMX\_30357* locus (Figure 5A). This locus was previously annotated as two separate protein coding genes (*KLMA\_30367* and *KLMA\_30368*) but we show that it is a single transcript as there is only evidence for one TSS and 1 PAS. Furthermore, although there was relatively uniform RNA-seq distribution along the length of this putative single transcript, the reading frame changed to the +1 position in the spacer region between the “*KLMA\_30367*” and “*KLMA\_30368*” CDS. This suggested to us that translation of “*KLMA\_30368*” CDS could be due to ribosomal frameshifting. Indeed, sequence analysis revealed the presence of a likely shift-prone pattern (GCG\_AGG\_C) at the site where the ribosome density in the +1 frame increases (Supplementary figure 12). This particular heptamer sequence was previously shown to support +1 frameshifting in *S. cerevisiae* (Sundararajan *et al.* 1999) and may be described as a hybrid of Ty1 (CUU\_AGG\_C) and Ty3 (GCG\_AGU\_U) +1 frameshift heptamers described in *S. cerevisiae* (Clare, Belcourt and Farabaugh 1988; Farabaugh, Zhao and Vimaladithan 1993). Frameshifting prone patterns are known to be underrepresented in coding sequences as they reduce processivity of translation (Shah *et al.* 2002; Gurvich *et al.* 2003). We analysed the occurrence of all in-frame heptamers in *K. marxianus* and found that this heptamer occurs far more rarely than what could be predicted based on its codon composition (see methods and Supplementary Figure 12) and is among the ~2% of the rarest heptamers present in coding regions (Supplementary Figure 14). It has been suggested that severe imbalance between availability of tRNAs for the A-site codons in 0 (AGG) and +1 (GGC) frames in Ty1 frameshifting site is responsible for its high efficiency in *S. cerevisiae* (Baranov, Gesteland and Atkins 2004). Based on the assumption that tRNA copy number correlates with tRNA abundance (Percudani, Pavesi and Ottonello 1997), we generated a table that estimated the relative abundance of each tRNA in *K. marxianus* (see methods and Supplementary table 9). We find that the ratio of tRNA<sup>Arg</sup><sub>CCU</sub>

decoding AGG (0 frame) to tRNA<sup>Gly</sup><sub>GCC</sub> decoding GGC (+1 frame) is 1:9, which is consistent with a hypothesis that a tRNA imbalance may be the driver of this frameshift. The expected full-length product with +1 frameshifting would be a protein 857 amino acids in length. We searched for homologs of this full length protein and discovered this gene can be separated to three regions of interest (Figure 5B). The zero-frame ORF is a homolog of *S. cerevisiae* YLR257W, a gene of unknown function. The +1 frame product contains a midasin/AAA ATPase domain except for the C-terminus, which is highly similar to the C-terminus of *S. cerevisiae* Aip5p. Aip5p is part of a multiprotein polarisome complex which catalyses the formation of actin cables for polarized cell growth during budding (Glomb, Bareis and Johnsson 2019; Xie *et al.* 2019). Interestingly, the C-terminus of Aip5p has been shown to be responsible for this activity (Figure 5B). Thus, it is possible that this gene is translated into two proteoforms with distinct functions. We note *S. cerevisiae* Aip5 also contains a midasin/AAA ATPase domain upstream of the C-terminus domain, like the +1 product.



**Figure 5.** Frameshifting at the KLMX\_30357 locus. A displays the region including the newly annotated KLMX\_30357 0-frame (red) and +1-frame CDS (green) (originally annotated separately as KLMA\_30367 and KLMA\_30368). B. Schematic of the KLMX\_30357 locus and similarity of full-length frameshift products to known genes and superfamily. Percentages refer to the identity with known homologs in *S. cerevisiae*. See Figure 2 for legend explanation of the multi-omics plots.

### Detection of novel protein coding genes and improvement of existing genome annotation

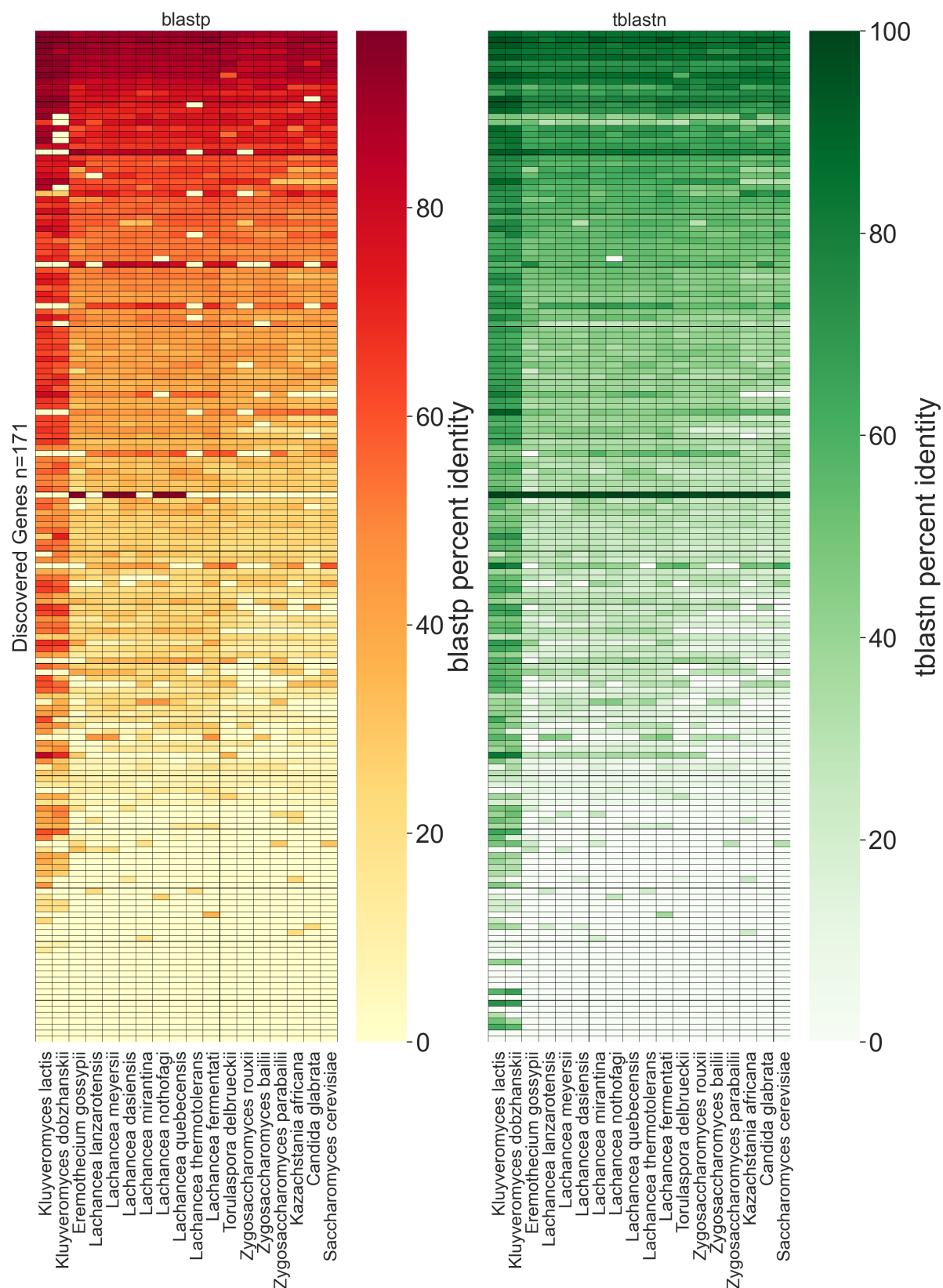
During analysis of the ribosome profiling data, it was apparent that there were significant numbers of RPFs aligning to the regions of the DMKU3-1042 reference genome that had not been annotated as protein coding. Some of these were found to be caused by a ~40 Kb annotation gap on chromosome 1, which excluded 20 genes, but there were multiple others not explained in this way. Therefore, we decided to use our ribosome profiling data to improve the *K. marxianus* DMKU3-1042 genome annotation by creating a semi-supervised pipeline to detect translated protein coding regions that were not annotated as genes in the

reference genome. With this approach, we discovered 171 unannotated candidate genes in the DMKU3-1042 genome and further investigated these putative genes by generating sub-codon ribosome profiles for each transcript using Trips-Viz (Kiniry *et al.* 2019, 2021). These plots were explored manually for the consistency of ribosome profiling density and triplet periodicity in case artefacts had been introduced by the computational pipeline (see Supplementary Figure 15 for examples). We also generated a table displaying the periodicity score and the number of RPFs per open reading frame for each of these candidate genes (Supplementary table 10). In all cases, the patterns are consistent with protein encoding genes. We then analysed the amino acid sequence to explore this novel gene set. Each of the 171 CDSs was conceptually translated and the resulting protein sequences were queried against the NCBI non-redundant protein database with the BLASTP tool (Altschul *et al.* 1990). Finally, we interrogated the annotated genomes of 18 yeast species within the budding yeast sub-phylum (*Saccharomycotina*) using both BLASTP and TBLASTN (Figure 6). BLASTP can be used to identify homologs in existing annotations, while TBLASTN allows for identification of homologs in the corresponding genomes irrespective of the completeness of its annotation. Thus, an existence of a high scoring hit in a TBLASTN search that is absent in a BLASTP search would indicate the presence of an ortholog that has not been annotated. This analysis included species in the *Kluyveromyces* genus, other species in the *Kluyveromyces* / *Lachancea* / *Eremothecium* (KLE) clade, representatives from the *Zygosaccharomyces* / *Torulaspora* (Z/T) clade, and three post whole-genome duplication (WGD) species, *Candida glabrata*, *Kazachstania africana* and *S. cerevisiae* (Shen *et al.* 2018). A larger number of *Lachancea* species were included as this is the most closely related genus to *Kluyveromyces*. There were substantial differences between the results of the BLASTP and TBLASTN analyses, indicating that quite a number of protein coding genes are missing from the genome annotations of some species. For the majority of the 171 genes, we found homologous proteins in most or all of the 17 other yeast species from the budding yeast subphylum that we included. In some cases, these genes encoded orthologs of proteins with known or easily predicted functions; for example, transcription factors, ribosomal proteins, a sugar transporter, a redox protein, heat shock proteins and others (Supplementary table 11). The addition of these protein coding genes increased the total number of annotated protein coding genes from 4,952 to 5,118 in the DMKU3-1042 genome (see Supplementary table 12 for comparison of published *K. marxianus* genomes). Ultimately, of the 171 newly identified translated genes, only 10 genes are specific to *K. marxianus* and a further 16 genes appear



specific to the *Kluyveromyces* genus. Analysis of these 26 proteins with the PFAM superfamily database failed to identify any known domains that would give clues as to function.

In addition to discovering previously non-annotated genes, we found and corrected 120 incorrectly annotated genes. A full list and breakdown of these gene corrections is provided in Supplementary table 13. The start codon was reannotated for 69 genes as an unequivocal density of in-frame translating ribosomes could be observed starting either upstream or downstream of the previously annotated start codon (for example see Supplementary Figure 16 & Supplementary table 14 for description of all start codon corrections). Multiple corrections were also made in nuclear-encoded intron-containing genes. The total number of annotated spliced genes increased by 13 to 183, as 28 of the newly-annotated genes contain introns but 15 genes previously annotated as containing introns did not, in fact, contain introns. In most of these cases, translation initiated from an in-frame AUG in the reported “second” exon and there was no evidence of the “first” exon being translated. One example of a spliced gene where the splicing was missed is *QCR9*, which encodes a subunit of Complex III in the mitochondrion electron transport chain (Supplementary Figure 17). In 20 cases where splicing was reported, the coordinates of intron-exon boundaries were incorrect and needed to be amended, for example, the *NUP60* locus (Supplementary Figure 18). A full list of spliced genes in *K. marxianus* is provided in Supplementary table 15. The genes encoding the ribosomal proteins Rpl7, Rps9 and Rpl36B were annotated on the wrong strand and this was corrected (see Supplementary Figure 19 for example). Finally, we correctly annotated the +1 frameshifting genes *OAZ1* and *ABP140* to include both the 0 frame and +1 frame as original annotations reported a single ORF (for example see Supplementary Figure 20), and we also included as a correction the -1 frameshift gene *KAT1* (Rajaei *et al.* 2014), originally annotated as two separate protein-coding genes. We deposited our updated annotation to our developed RiboSeq.Org resources and users may now freely explore our ribosome profiling and transcriptomic data on the GWIPS-viz genome browser relative to the original or updated genome annotation. In addition, TSS and PolyA coverage tracks have also been added to GWIPS-viz (<https://gwips.ucc.ie/>).



**Figure 6. Analysis of orthologous groups for 171 newly identified genes.** Left is a heatmap of % amino acid identity for the BLASTP hits obtaining for a search against proteins annotated in other genomes. Right is a heatmap of % amino acid identity for the TBLASTN hits obtained for a search against genomic sequences. Each row in both tables corresponds to the same gene for comparison.

## Discussion

A robust accurately annotated genome is an important tool in modern yeast genetics and while advances in sequencing technology make it easy to generate a genome sequence, databases are replete with incomplete information and annotation errors. This is particularly the case when considering non-model yeasts since annotation is generally based on using *S. cerevisiae* as the reference. This can lead to an accumulation of errors because of the underlying assumption that the new genome will not deviate substantially from the supposed reference genome. We demonstrate how a genome annotation can be improved by including experimental data that measure transcription and translation. This multi-omics approach enabled us to significantly ameliorate the *K. marxianus* annotation through the addition of genes, correction of splicing events, and identification of correct transcription and translation start sites. The *K. marxianus* genome annotation presented here is the most accurate and complete genome annotation within the budding yeasts, apart, perhaps, from *S. cerevisiae*. This annotation can serve as a reference for the annotation of other *Kluyveromyces marxianus* strains and closely related species and the methodology could be applied to other non-traditional yeasts. It may also be valuable for a number of previous studies in *K. marxianus* that used RNA-Seq to study responses to various stresses and stimuli. Over 300 new or corrected gene annotations are now present and, given that species-specific genes are predicted to be important for stress response and niche adaptation, it may be worthwhile to reanalyse RNA-Seq data using this new annotation. Similarly, for comparative genomic studies, proper gene level annotation is crucial to avoid errors and thus this new reference will be valuable.

A major innovation of this study was the integration of different omic methods. The development of a method for ribosome profiling in *K. marxianus* is recent (Fenton *et al.* 2022) and this work marked its first use. In combination with TSS data and 3' biased RNA-seq (due to poly(A) selection), it was possible to reveal mRNA boundaries and the presence of mRNA isoforms on a genome-wide scale. This provided information on both the transcriptional and translational landscape, which aided in deciphering events such as N-terminal extensions and uORFs through the use of multiomics visualization of a locus of interest. Our computational pipeline showed the diversity in the proteome with multiomics data, identifying the expression of alternative proteoforms and the translation of many short ORFs. With the annotation of mRNA boundaries, we were able to decipher whether alternative proteoforms are generated by alternative transcription due to the use of different

transcription start sites within the promoter (such as with *ADO1* and *FUM1*) or alternative translation initiation due to leaky scanning (*FOL1*).

It will be very interesting in the future to compare the complex and varied genomic diversity in *K. marxianus* with that of *S. cerevisiae* as they represent two yeast lineages with a very different evolutionary history. One, *S. cerevisiae*, arose from a hybridisation between parents from the KLE and ZT clades (Wolfe and Shields 1997; Marcet-Houben and Gabaldón 2015), and thus has a duplicated genome with numerous ohnologs. In contrast, *K. marxianus* represents pre-WGD yeasts and thus may be more representative in many regards of the ancestral state. Already from our limited analysis, we identified multiple examples where the mechanism used to generate proteoforms, or to regulate levels of proteins, is different between the two yeast species. In some cases, the *S. cerevisiae* data may be incomplete and some previous assumptions incorrect, but in others, e.g for regulating *EST3*, it is clear that different mechanisms are used. Many hypotheses can be generated from our dataset and are likely to be a fertile area of study to understand evolution and niche adaption in the Saccharomycetales.

Despite the wealth of data and novel insights generated in this work, it must be acknowledged that the overall number of conditions tested was quite limited. Thus, although we managed to detect expression of the vast majority of protein coding genes, it is possible that some aspects of regulation were missed because they only arise in some specific condition that we did not use. This is, of course, the case with all previous genome-wide studies of gene expression in other yeasts as well. Nonetheless, it is accepted the data are unlikely to reveal the full spectrum of expressed RNAs and synthesized proteins of the *K. marxianus* genome. The application of multiomics techniques for mapping the ends of transcripts and positions of ribosomes under different conditions will likely reveal additional features of protein-coding organisation in the *K. marxianus* genome.

## Materials and Methods

### RNA-Seq and Ribosome Profiling

RNA-Seq and ribosome profiling was carried out as in (Fenton *et al.* 2022). Splice junctions were identified for novel intron containing genes and genes which required splice site corrections using the splice aware STAR RNA-seq aligner (Dobin *et al.* 2013). In the ribosome profiling analysis, RPFs that failed to align to the original CDS regions of DMKU3-1042 annotation were aligned to the reference genome. These alignments were then split into windows using Bedtools (Quinlan and Hall 2010). Windows were ranked based on the number of alignments and the top candidates were visually assessed using a genome browser (GWIPS-viz (Michel *et al.* 2014)) where we created a database for *K. marxianus* DMKU-1042 genome. For the BLASTP and TBLASTN heatmaps, custom databases containing annotated protein sequences and genome assemblies were created for use with BLASTP and TBLASTN, respectively (Altschul *et al.* 1990). For BLASTP and TBLASTN, the following parameters were specified, -seg no, -threshold 11, -max\_hsps 1 and -outfmt 6. An e value filter of  $\leq 0.01$  was applied to blast results.

### TSS-Seq

In order to precisely characterise Transcription Start Sites, publicly available TSS-seq data originating from *K. marxianus* DMKU3-1042 was used (Lertwattanasakul *et al.* 2015). Data were downloaded from the NCBI SRA repository, adaptor removal and quality trimming was performed with cutadapt, followed with rRNA removal and genome alignment with bowtie. While doing this analysis, we noticed that accession numbers of raw data deposited in SRA do not match expression profiles of conditions discussed in the original paper (Lertwattanasakul *et al.* 2015) and were evidently mislabelled during sequence deposition. For our analysis, we reassigned the samples to the correct condition, detailed in Supplementary material. Resulting alignments were used for transcriptional units (Transcription Start Region — TSR) detection by clustering reads with Bioconductor package CAGEr (Haberle *et al.* 2015). TSRs were subsequently assigned to the nearest coding sequence (CDS) with a minimum relative expression cut-off of 0.05 and 1 TPM was applied to filter out lowly expressed TSRs or unreliable clusters (see Supplementary Figure 5).

### Identification of PAS

For identification of polyadenylation sites, we used our own polyA-enriched RNA-seq data (Fenton *et al.* 2022). Reads that aligned to the genome were discarded as these are sequences

that do not contain polyA tails. From the remaining reads, all trailing A nucleotides were trimmed from 3' ends of reads and aligned once again to the genome revealing PAS. Aligned reads were processed analogously to the TSS-seq reads: clustering has been performed with CAGEr and followed with assignment to nearest CDS with minimum relative expression cut-off of 0.05 and 1 TPM. Remaining clusters were assumed as *bona-fide* polyadenylation sites (PAS, see Supplementary Figure 6).

### Multiomics plots

Briefly, BAM files were processed with ORFik bioconductor package to generate P-site ribo-seq profiles and RNA-seq coverage profiles (Tjeldnes *et al.* 2021). In the process of mining the data, RiboCrypt: R package NGS data visualization tool was developed. It takes use of ORFik data management and processing and ggplot2 combined with plotly for data display. RiboCrypt GitHub repository is available at <https://github.com/m-swirski/RiboCrypt>. The profiles are characterised by very sharp peaks, making profiles less-readable when zoomed-out. Thus a sliding window mean of a profile was used to decrease resolution and increase clarity of a picture. To display coverage we employed the stacked method to avoid blurring the plot by area overlapping.

### Non-canonical translation detection

All possible ORFs starting from any one of the near-cognate codons (differing from AUG by one nucleotide, including AUG) in the *K. marxianus* genome and transcriptome were found with ORFik package (Tjeldnes *et al.* 2021). Naturally, it resulted in finding multiple ORFs sharing a stop codon. Subsequently, a P-site profile was generated for the longest ORF for each stop codon and a set of parameters was calculated for all nested ORFs. P-site score: % of in-frame reads, read count – given as reads per kilobase (RPK), in-frame coverage fold-change between a region between Nth and Nth+1 start codon and first to Nth-1 start codon. Unique P-site score: P-site score calculated for a region between Nth and Nth+1 start codon. Additionally, for all ORFs longer than 20 codons a MTS was calculated with MitoFates software. For potential isoforms of annotated proteins (novel or annotated before) a difference in MTS prediction between annotated start codon and potential aTIS was calculated to assess possibility of initiation dependent MTS translation.

### tRNA Copy Numbers and Heptamer Frequency Analysis

tRNA copy numbers for the reference genome (DMKU3-1042) were determined with tRNA scan-SE (Chan and Lowe 2019). The following formula was used for all heptamers found in CDS regions, where B is the +1 nucleotide (7<sup>th</sup> base of heptamer).

$$N\{total\} * N\{total\} * N\{heptamer\} / N\{codon1\} * N\{codon2\} * N\{BXX\}$$

### **Data Access**

Ribosome profiling, TSS, PAS and RNA-seq data have been deposited to GWIPS-viz, as well as the original and new annotation track. Ribosome profiling and RNA-seq have been deposited to Trips-Viz. Ribosome profiling and RNA-seq datasets have been deposited to the Sequence Read Archive under the project accession number PRJEB45612. Our updated genome annotation and other relevant files are available at

<https://doi.org/10.5281/zenodo.6378617>.

## References

- Adiconis X, Haber AL, Simmons SK *et al.* Comprehensive comparative analysis of 5'-end RNA-sequencing methods. *Nat Methods* 2018;**15**:505–11.
- Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
- Alva TR, Riera M, Chartron JW. Translational landscape and protein biogenesis demands of the early secretory pathway in *Komagataella phaffii*. *Microb Cell Fact* 2021;**20**:19.
- Arribere JA, Gilbert W V. Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res* 2013;**23**:977–87.
- Atkins JF, Loughran G, Bhatt PR *et al.* Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. *Nucleic Acids Res* 2016;**44**:7007–78.
- Baptista M, Cunha JT, Domingues L. Establishment of *Kluyveromyces marxianus* as a Microbial Cell Factory for Lignocellulosic Processes: Production of High Value Furan Derivatives. *J Fungi* 2021;**7**, DOI: 10.3390/jof7121047.
- Baranov P V, Gesteland RF, Atkins JF. P-site tRNA is a crucial initiator of ribosomal frameshifting. *RNA* 2004;**10**:221–30.
- Brar GA, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* 2015;**16**:651–64.
- Cernak P, Estrela R, Poddar S *et al.* Engineering *Kluyveromyces marxianus* as a Robust Synthetic Biology Platform Host. *MBio* 2018;**9**, DOI: 10.1128/mBio.01410-18.
- Chan PP, Lowe TM. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. *Methods Mol Biol* 2019;**1962**:1–14.
- Chen Y, Beck A, Davenport C *et al.* Characterization of TRZ1, a yeast homolog of the human candidate prostate cancer susceptibility gene ELAC2 encoding tRNase Z. *BMC Mol Biol* 2005;**6**:12.
- Clare JJ, Belcourt M, Farabaugh PJ. Efficient translational frameshifting occurs within a conserved sequence of the overlap between the two genes of a yeast Ty1 transposon. *Proc Natl Acad Sci U S A* 1988;**85**:6816–20.
- Dekker WJC, Ortiz-Merino RA, Kaljouw A *et al.* Engineering the thermotolerant industrial



- yeast *Kluyveromyces marxianus* for anaerobic growth. *bioRxiv* 2021:2021.01.07.425723.
- Dobin A, Davis CA, Schlesinger F *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21.
- Donzella L, Varela JA, Sousa MJ *et al.* Identification of novel pentose transporters in *Kluyveromyces marxianus* using a new screening platform. *FEMS Yeast Res* 2021;**21**, DOI: 10.1093/femsyr/foab026.
- Doughty TW, Domenzain I, Millan-Oropeza A *et al.* Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts. *Nat Commun* 2020;**11**:2144.
- Duncan CDS, Mata J. The translational landscape of fission-yeast meiosis and sporulation. *Nat Struct Mol Biol* 2014;**21**:641–7.
- Eisenberg AR, Higdon AL, Hollerer I *et al.* Translation Initiation Site Profiling Reveals Widespread Synthesis of Non-AUG-Initiated Protein Isoforms in Yeast. *Cell Syst* 2020;**11**:145-160.e5.
- Farabaugh PJ, Kramer E, Vallabhaneni H *et al.* Evolution of +1 programmed frameshifting signals and frameshift-regulating tRNAs in the order Saccharomycetales. *J Mol Evol* 2006;**63**:545–61.
- Farabaugh PJ, Zhao H, Vimaladithan A. A novel programmed frameshift expresses the POL3 gene of retrotransposon Ty3 of yeast: frameshifting without tRNA slippage. *Cell* 1993;**74**:93–103.
- Fasoli G, Tofalo R, Lanciotti R *et al.* Chromosome arrangement, differentiation of growth kinetics and volatile molecule profiles in *Kluyveromyces marxianus* strains from Italian cheeses. *Int J Food Microbiol* 2015;**214**:151–8.
- Fenton DA, Kiniry SJ, Yordanova MM *et al.* Development of a Ribosome Profiling Protocol to Study Translation in the yeast *Kluyveromyces marxianus*. *bioRxiv* 2022:2022.02.06.478964.
- Fonseca GG, Heinzle E, Wittmann C *et al.* The yeast *Kluyveromyces marxianus* and its biotechnological potential. *Appl Microbiol Biotechnol* 2008;**79**:339–54.
- Franceschetti M, Hanfrey C, Scaramagli S *et al.* Characterization of monocot and dicot plant S-adenosyl-l-methionine decarboxylase gene families including identification in the

- mRNA of a highly conserved pair of upstream overlapping open reading frames. *Biochem J* 2001;**353**:403–9.
- Gaba A, Jacobson A, Sachs MS. Ribosome Occupancy of the Yeast CPA1 Upstream Open Reading Frame Termination Codon Modulates Nonsense-Mediated mRNA Decay. *Mol Cell* 2005;**20**:449–60.
- Gaba A, Wang Z, Krishnamoorthy T *et al.* Physical evidence for distinct mechanisms of translational control by upstream open reading frames. *EMBO J* 2001;**20**:6453–63.
- Glomb O, Bareis L, Johnsson N. YFR016c/Aip5 is part of an actin nucleation complex in yeast. *Biol Open* 2019;**8**, DOI: 10.1242/bio.044024.
- Gurvich OL, Baranov P V, Zhou J *et al.* Sequences that direct significant levels of frameshifting are frequent in coding regions of Escherichia coli. *EMBO J* 2003;**22**:5941–50.
- Ha-Tran DM, Nguyen TT, Huang C-C. Kluyveromyces marxianus: Current State of Omics Studies, Strain Improvement Strategy and Potential Industrial Implementation. *Ferment* 2020;**6**, DOI: 10.3390/fermentation6040124.
- Haberle V, Forrest ARR, Hayashizaki Y *et al.* CAGER: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res* 2015;**43**:e51–e51.
- Hardy S, Kostantin E, Wang SJ *et al.* Magnesium-sensitive upstream ORF controls PRL phosphatase expression to mediate energy metabolism. *Proc Natl Acad Sci* 2019:201815361.
- Herrmann JM. Putting a break on protein translocation: metabolic regulation of mitochondrial protein import. *Mol Microbiol* 2009;**72**:275–8.
- Hinnebusch AG. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol* 2005;**59**:407–50.
- Hood HM, Neafsey DE, Galagan J *et al.* Evolutionary Roles of Upstream Open Reading Frames in Mediating Gene Regulation in Fungi. *Annu Rev Microbiol* 2009;**63**:385–409.
- Ingolia NT, Ghaemmaghami S, Newman JRS *et al.* Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 2009;**324**:218–23.

- Inokuma K, Ishii J, Hara KY *et al.* Complete Genome Sequence of *Kluyveromyces marxianus* NBRC1777, a Nonconventional Thermotolerant Yeast. *Genome Announc* 2015;**3**:e00389-15.
- Ivanov IP, Atkins JF. Ribosomal frameshifting in decoding antizyme mRNAs from yeast and protists to humans: close to 300 cases reveal remarkable diversity despite underlying conservation. *Nucleic Acids Res* 2007;**35**:1842–58.
- Ivanov IP, Gesteland RF, Atkins JF. Evolutionary specialization of recoding: frameshifting in the expression of *S. cerevisiae* antizyme mRNA is via an atypical antizyme shift site but is still +1. *RNA* 2006;**12**:332–7.
- Ivanov IP, Shin B-S, Loughran G *et al.* Polyamine Control of Translation Elongation Regulates Start Site Selection on Antizyme Inhibitor mRNA via Ribosome Queuing. *Mol Cell* 2018;**70**:254-264.e6.
- Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 2004;**338**:1027–36.
- Karim A, Gerliani N, Aïder M. *Kluyveromyces marxianus*: An emerging yeast cell factory for applications in food and biotechnology. *Int J Food Microbiol* 2020;**333**:108818.
- Kiniry SJ, Judge CE, Michel AM *et al.* Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. *Nucleic Acids Res* 2021;**49**:W662–70.
- Kiniry SJ, Michel AM, Baranov P V. Computational methods for ribosome profiling data analysis. *Wiley Interdiscip Rev RNA* 2020;**11**:e1577.
- Kiniry SJ, O'Connor PBF, Michel AM *et al.* Trips-Viz: a transcriptome browser for exploring Ribo-Seq data. *Nucleic Acids Res* 2019;**47**:D847–52.
- Knoshaug EP, Vidgren V, Magalhães F *et al.* Novel transporters from *Kluyveromyces marxianus* and *Pichia guilliermondii* expressed in *Saccharomyces cerevisiae* enable growth on l-arabinose and d-xylose. *Yeast* 2015;**32**:615–28.
- Lane MM, Morrissey JP. *Kluyveromyces marxianus*: A yeast emerging from its sister's shadow. *Fungal Biol Rev* 2010;**24**:17–26.
- Law GL, Raney A, Heusner C *et al.* Polyamine Regulation of Ribosome Pausing at the Upstream Open Reading Frame of S-Adenosylmethionine Decarboxylase\*. *J Biol Chem* 2001;**276**:38036–43.

- Lertwattanasakul N, Kosaka T, Hosoyama A *et al*. Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. *Biotechnol Biofuels* 2015;**8**:47.
- Liu Y, Nielsen J. Recent trends in metabolic engineering of microbial chemical factories. *Curr Opin Biotechnol* 2019;**60**:188–97.
- Marcet-Houben M, Gabaldón T. Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker’s Yeast Lineage. *PLOS Biol* 2015;**13**:e1002220.
- McGlinchy NJ, Ingolia NT. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* 2017;**126**:112–29.
- McManus CJ, May GE, Spealman P *et al*. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res* 2014;**24**:422–30.
- Michel AM, Baranov P V. Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley Interdiscip Rev RNA* 2013;**4**:473–90.
- Michel AM, Fox G, M Kiran A *et al*. GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res* 2014;**42**:D859–64.
- Michel AM, Kiniry SJ, O’Connor PBF *et al*. GWIPS-viz: 2018 update. *Nucleic Acids Res* 2018;**46**:D823–30.
- Mo W, Wang M, Zhan R *et al*. *Kluyveromyces marxianus* developing ethanol tolerance during adaptive evolution with significant improvements of multiple pathways. *Biotechnol Biofuels* 2019;**12**:63.
- Monteuuis G, Miścicka A, Świrski M *et al*. Non-canonical translation initiation in yeast generates a cryptic pool of mitochondrial proteins. *Nucleic Acids Res* 2019;**47**:5777–91.
- Montini N, Doughty TW, Domenzain I *et al*. Identification of a novel gene required for competitive growth at high temperature in the thermotolerant yeast *Kluyveromyces marxianus*. *Microbiology* 2022;**168**, DOI: 10.1099/mic.0.001148.
- Morrissey JP, Etschmann MMW, Schrader J *et al*. Cell factory applications of the yeast *Kluyveromyces marxianus* for the biotechnological production of natural flavour and fragrance molecules. *Yeast* 2015;**32**:3–16.
- Nagalakshmi U, Wang Z, Waern K *et al*. The Transcriptional Landscape of the Yeast

- Genome Defined by RNA Sequencing. *Science* (80- ) 2008;**320**:1344 LP – 1349.
- Nambu-Nishida Y, Nishida K, Hasunuma T *et al.* Development of a comprehensive set of tools for genome engineering in a cold- and thermo-tolerant *Kluyveromyces marxianus* yeast strain. *Sci Rep* 2017;**7**:8993.
- Namy O, Duchateau-Nguyen G, Hatin I *et al.* Identification of stop codon readthrough genes in *Saccharomyces cerevisiae*. *Nucleic Acids Res* 2003;**31**:2289–96.
- Natsoulis G, Hilger F, Fink GR. The HTS1 gene encodes both the cytoplasmic and mitochondrial histidine tRNA synthetases of *S. cerevisiae*. *Cell* 1986;**46**:235–43.
- Ortiz-Merino RA, Varela JA, Coughlan AY *et al.* Ploidy Variation in *Kluyveromyces marxianus* Separates Dairy and Non-dairy Isolates . *Front Genet* 2018;**9**:94.
- Ozsolak F, Kapranov P, Foissac S *et al.* Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation. *Cell* 2010;**143**:1018–29.
- Palanimurugan R, Scheel H, Hofmann K *et al.* Polyamines regulate their synthesis by inducing expression and blocking degradation of ODC antizyme. *EMBO J* 2004;**23**:4857–67.
- Patra P, Das M, Kundu P *et al.* Recent advances in systems and synthetic biology approaches for developing novel cell-factories in non-conventional yeasts. *Biotechnol Adv* 2021;**47**:107695.
- Percudani R, Pavesi A, Ottonello S. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. *J Mol Biol* 1997;**268**:322–30.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2.
- Rajaei N, Chiruvella KK, Lin F *et al.* Domesticated transposase Kat1 and its fossil imprints induce sexual differentiation in yeast. *Proc Natl Acad Sci* 2014;**111**:15491 LP – 15496.
- Rajkumar AS, Morrissey JP. Rational engineering of *Kluyveromyces marxianus* to create a chassis for the production of aromatic products. *Microb Cell Fact* 2020;**19**:207.
- Rajkumar AS, Morrissey JP. Protocols for marker-free gene knock-out and knock-down in *Kluyveromyces marxianus* using CRISPR/Cas9. *FEMS Yeast Res* 2022;**22**:foab067.
- Rajkumar AS, Varela JA, Juergens H *et al.* Biological Parts for *Kluyveromyces marxianus* Synthetic Biology . *Front Bioeng Biotechnol* 2019;**7**:97.

- Regev-Rudzki N, Battat E, Goldberg I *et al.* Dual localization of fumarase is dependent on the integrity of the glyoxylate shunt. *Mol Microbiol* 2009;**72**:297–306.
- Shah AA, Giddings MC, Parvaz JB *et al.* Computational identification of putative programmed translational frameshift sites. *Bioinformatics* 2002;**18**:1046–53.
- Sharma P, Wu J, Nilges BS *et al.* Humans and other commonly used model organisms are resistant to cycloheximide-mediated biases in ribosome profiling experiments. *Nat Commun* 2021;**12**:5094.
- Shen X-X, Opulente DA, Kominek J *et al.* Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* 2018;**175**:1533-1545.e20.
- Skowronek E, Grzechnik P, Späth B *et al.* tRNA 3' processing in yeast involves tRNase Z, Rex1, and Rrp6. *RNA* 2014;**20**:115–30.
- Smith JE, Alvarez-Dominguez JR, Kline N *et al.* Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. *Cell Rep* 2014;**7**:1858–66.
- Spealman P, Naik AW, May GE *et al.* Conserved non-AUG uORFs revealed by a novel regression analysis of ribosome profiling data. *Genome Res* 2018;**28**:214–22.
- Sundararajan A, Michaud WA, Qian Q *et al.* Near-Cognate Peptidyl-tRNAs Promote +1 Programmed Translational Frameshifting in Yeast. *Mol Cell* 1999;**4**:1005–15.
- Tanaka M, Sotta N, Yamazumi Y *et al.* The Minimum Open Reading Frame, AUG-Stop, Induces Boron-Dependent Ribosome Stalling and mRNA Degradation. *Plant Cell* 2016;**28**:2830–49.
- Tang H-L, Yeh L-S, Chen N-K *et al.* Translation of a yeast mitochondrial tRNA synthetase initiated at redundant non-AUG codons. *J Biol Chem* 2004;**279**:49656–63.
- Terenin IM, Akulich KA, Andreev DE *et al.* Sliding of a 43S ribosomal complex from the recognized AUG codon triggered by a delay in eIF2-bound GTP hydrolysis. *Nucleic Acids Res* 2016;**44**:1882–93.
- Tjeldnes H, Labun K, Cleuren YT *et al.* ORFik: a comprehensive R toolkit for the analysis of translation. *bioRxiv* 2021:2021.01.16.426936.
- Varela JA, Gethins L, Stanton C *et al.* Applications of *Kluyveromyces marxianus* in Biotechnology BT - Yeast Diversity in Human Welfare. In: Satyanarayana T, Kunze G

- (eds.). Singapore: Springer Singapore, 2017, 439–53.
- Varela JA, Puricelli M, Ortiz-Merino RA *et al.* Origin of Lactose Fermentation in *Kluyveromyces lactis* by Interspecies Transfer of a Neo-functionalized Gene Cluster during Domestication. *Curr Biol* 2019;**29**:4284-4290.e2.
- Vindu A, Shin B-S, Choi K *et al.* Translational autoregulation of the *S. cerevisiae* high-affinity polyamine transporter Hol1. *Mol Cell* 2021, DOI: 10.1016/j.molcel.2021.07.020.
- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;**10**:57–63.
- Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 1997;**387**:708–13.
- Wu M, Tzagoloff A. Mitochondrial and cytoplasmic fumarases in *Saccharomyces cerevisiae* are encoded by a single nuclear gene FUM1. *J Biol Chem* 1987;**262**:12275–82.
- Xie Y, Sun J, Han X *et al.* Polarisome scaffold Spa2-mediated macromolecular condensation of Aip5 for actin polymerization. *Nat Commun* 2019;**10**:5078.
- Yu F, Zhang Y, Cheng C *et al.* Poly(A)-seq: A method for direct sequencing and analysis of the transcriptomic poly(A)-tails. *PLoS One* 2020;**15**:e0234696.

## Supplementary Material

Supplementary figure 1 – TSS and PAS isoforms per gene

Supplementary figure 2 – TSS data

Supplementary figure 3 – PAS data

Supplementary figure 4 – ORF calling scores

Supplementary figure 5 – Visualization of alternative proteoforms

Supplementary figure 6 – *FOLI* multiomics plot (column)

Supplementary figure 7 – Phobius predictions for *ADO1* isoforms

Supplementary figure 8 – ISC1 Trips-Viz Plot

Supplementary figure 9 – ISC1 sequence analysis

Supplementary figure 10 – Multiomics plot of GCN4.

Supplementary figure 11 – TRM112 and DIM1 antisense GWIPS-VIZ

Supplementary figure 12 – +1 frameshift multiomics column plot

Supplementary figure 13 – CDS heptamer scatterplot

Supplementary figure 14 – Log2 heptamer scatterplot

Supplementary figure 15 – single transcript profile examples (RPL7 and QCR10)

Supplementary figure 16 – GWIPS-viz example of incorrect start codon assignment (CWH41)

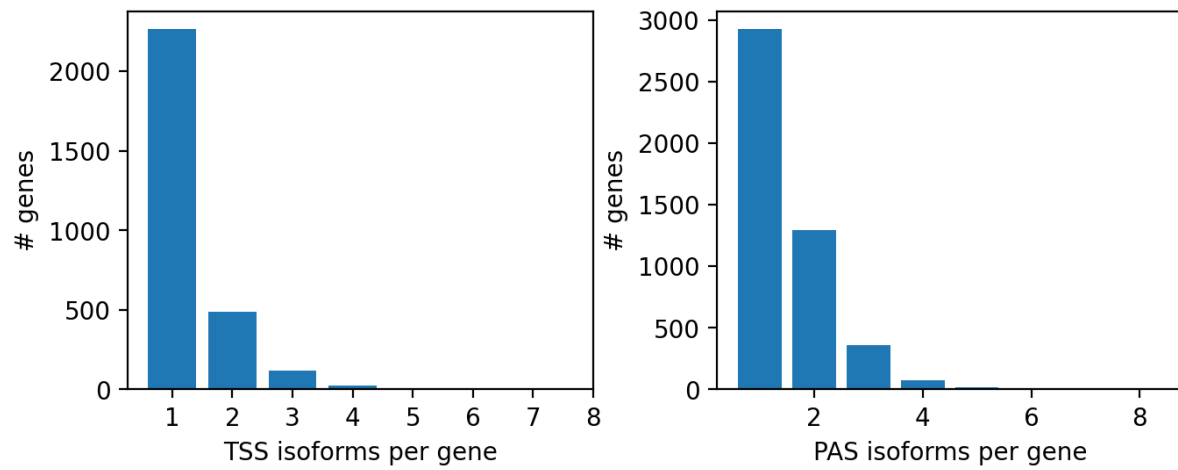
Supplementary figure 17 – GWIPS-viz example of incorrect splicing annotation (QCR9)

Supplementary figure 18 – GWIPS-viz example of incorrect splicing annotation (NUP60)

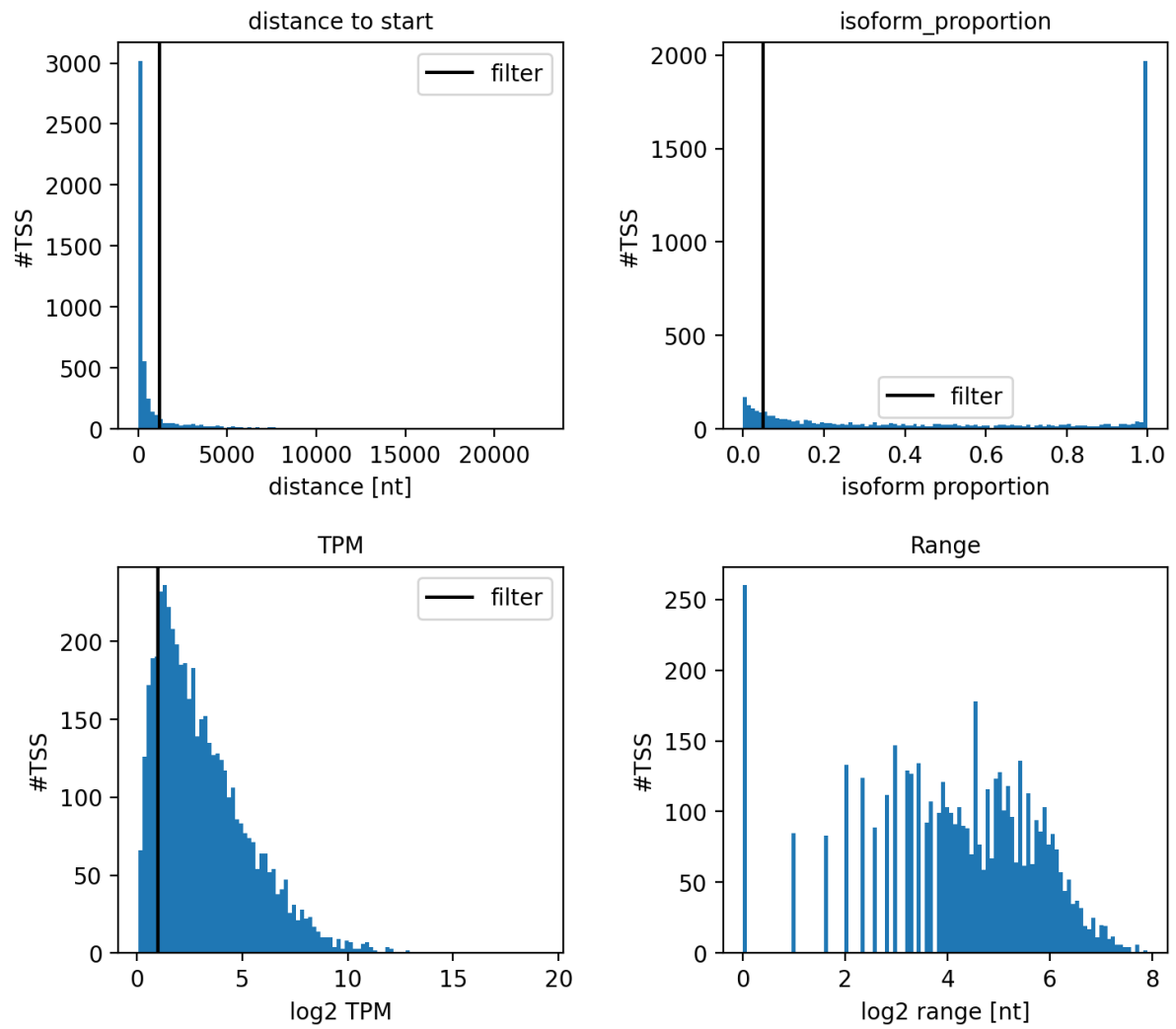
Supplementary figure 19 – GWIPS-viz example of incorrect strand annotation (RPS7)

Supplementary figure 20 – GWIPS-viz example of incorrect CDS annotation (PPT2 and ABP140)

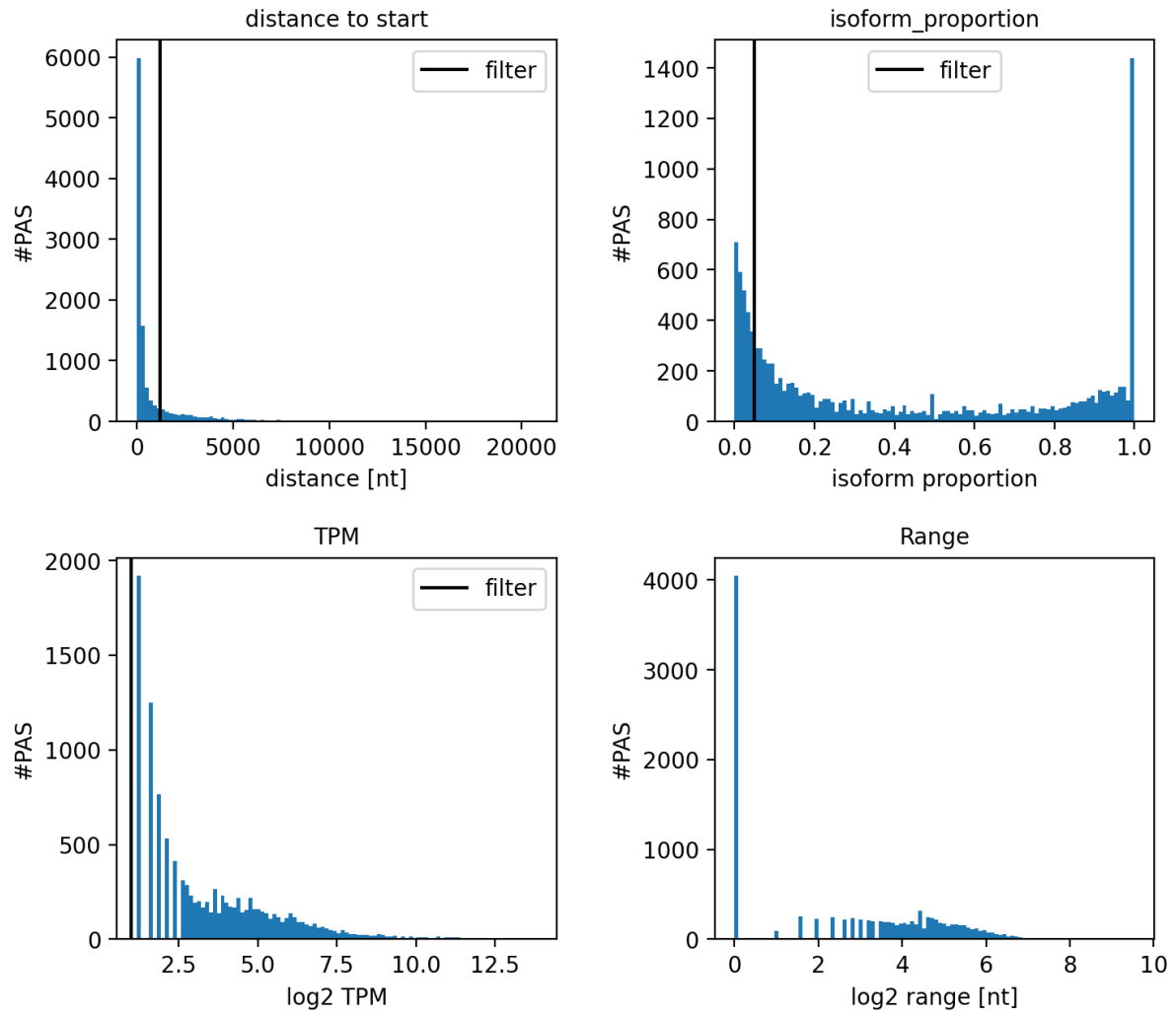




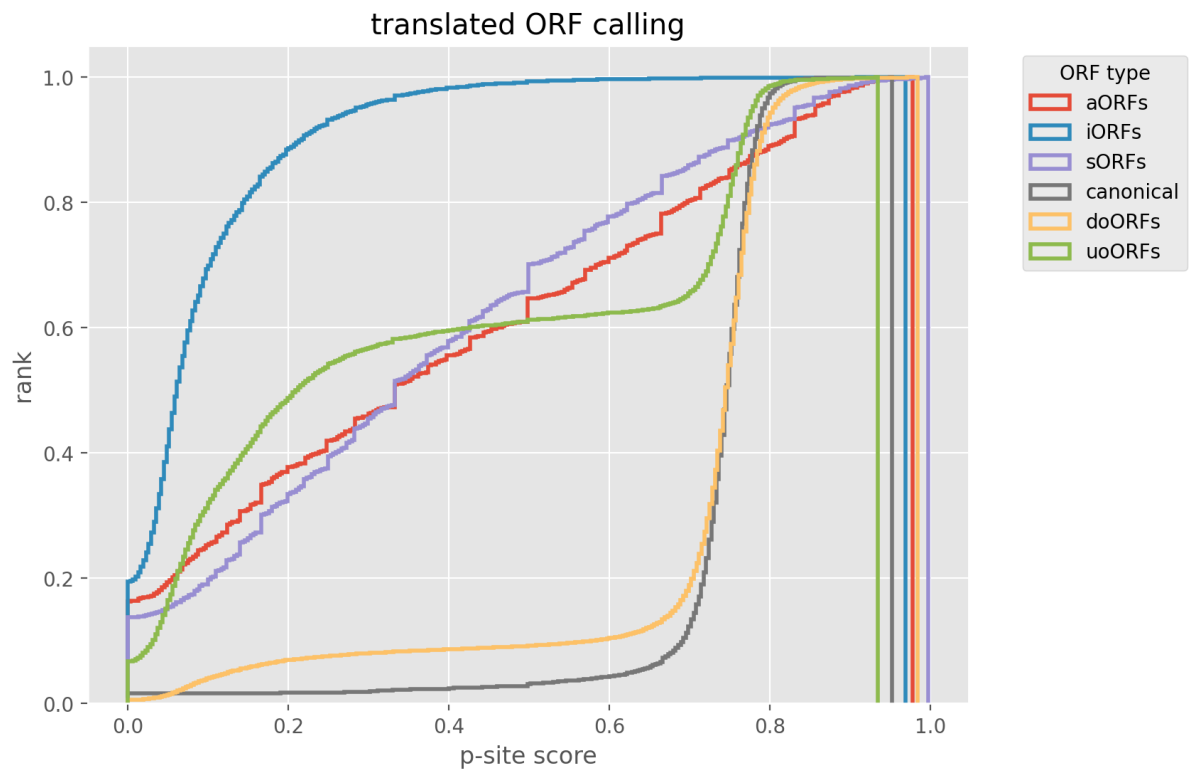
**Supplementary Figure 1.** Number of TSS isoforms per gene (left) and number of PAS isoforms (right).



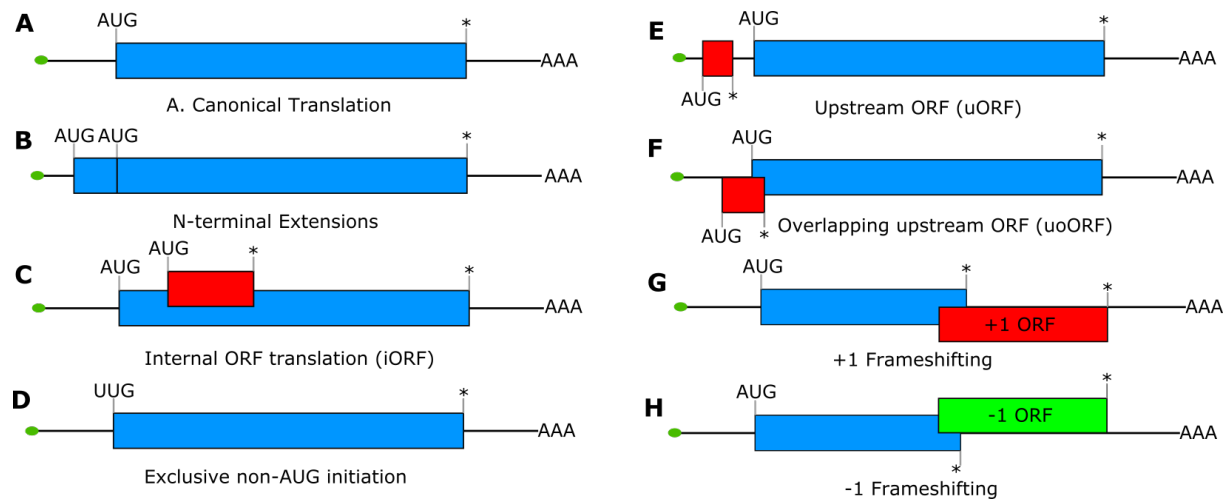
**Supplementary Figure 2.** Analysis of TSS data. Vertical black lines represent filters used in filtering of TSS data.



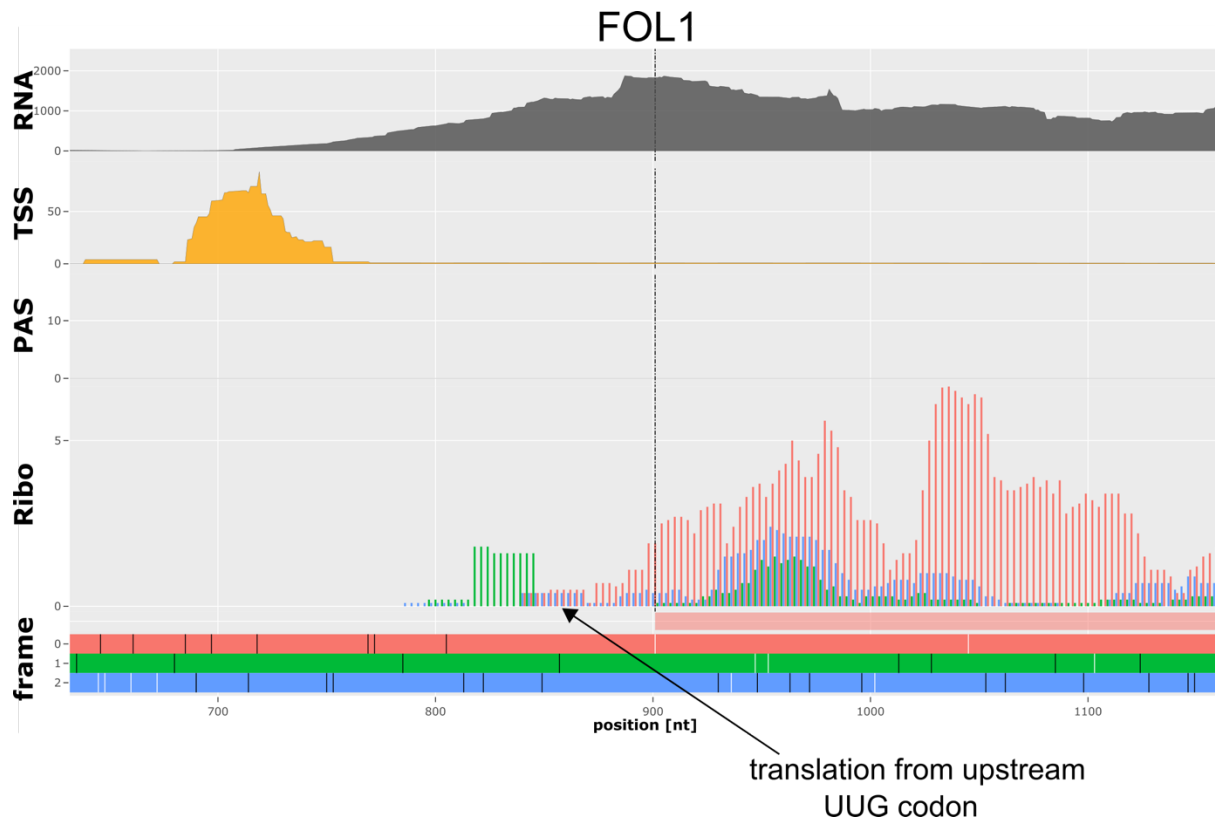
**Supplementary Figure 3.** Analysis of PAS data. Vertical black lines represent filters used in filtering of PAS data.



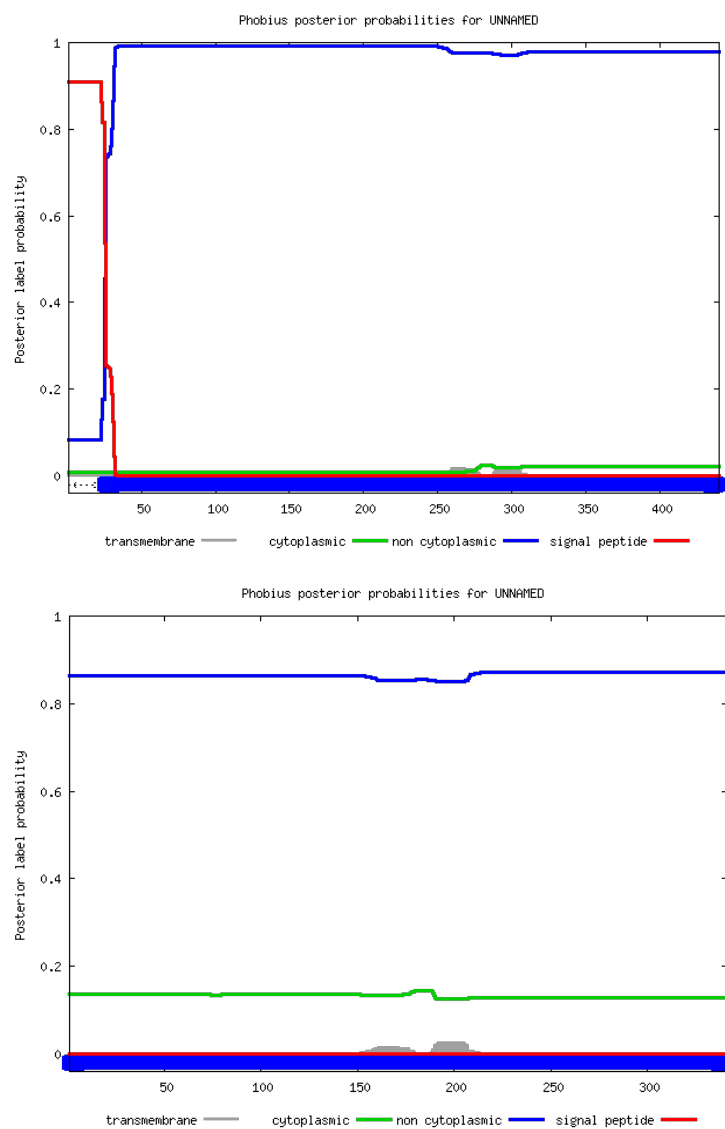
**Supplementary figure 4.** Cumulative histogram of called ORFs by class. X-axis represents ribosome profiling, p-site score and Y-axis represents rank. Note annotated protein-coding genes (canonical) as control.



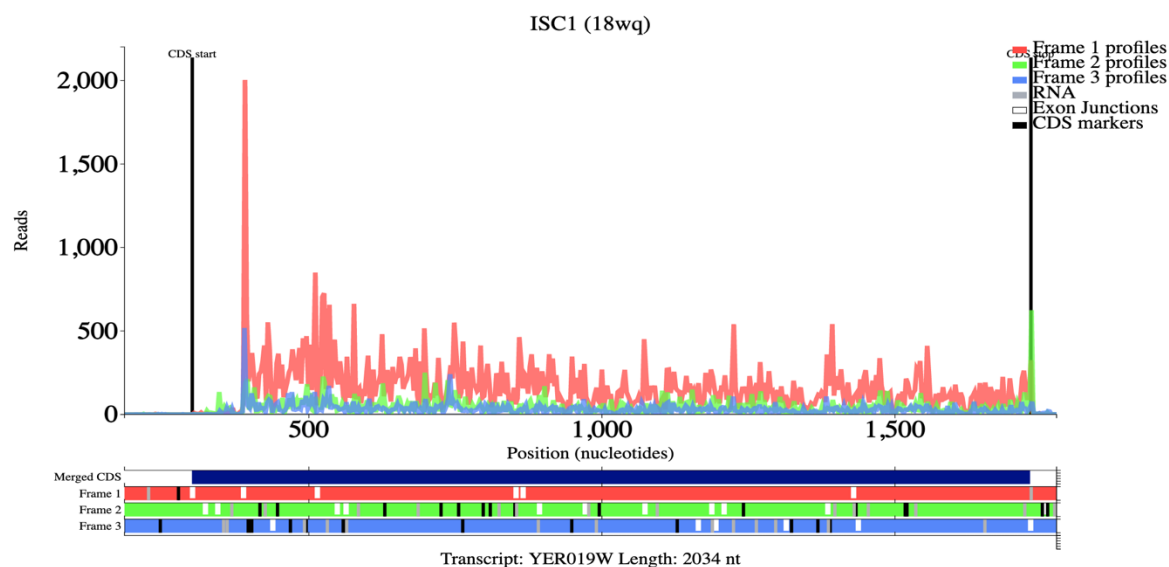
**Supplementary figure 5.** Examples of alternative translation of mRNAs. The 0, -1 and +1 reading frames are coloured at blue, green and red, respectively. For visualization purposes, examples of iORF, uORF and uoORF are considered as occurring in +1 frame.



**Supplementary Figure 6.** Multiomics plot of the *FOL1* 5' mRNA. Note the arrow pointing to upstream in-frame translation (red columns), which creates N-terminally extended proteoform. Translation of a non-AUG upstream open reading frame (uORF) is also present in the +1 frame, represented in green.

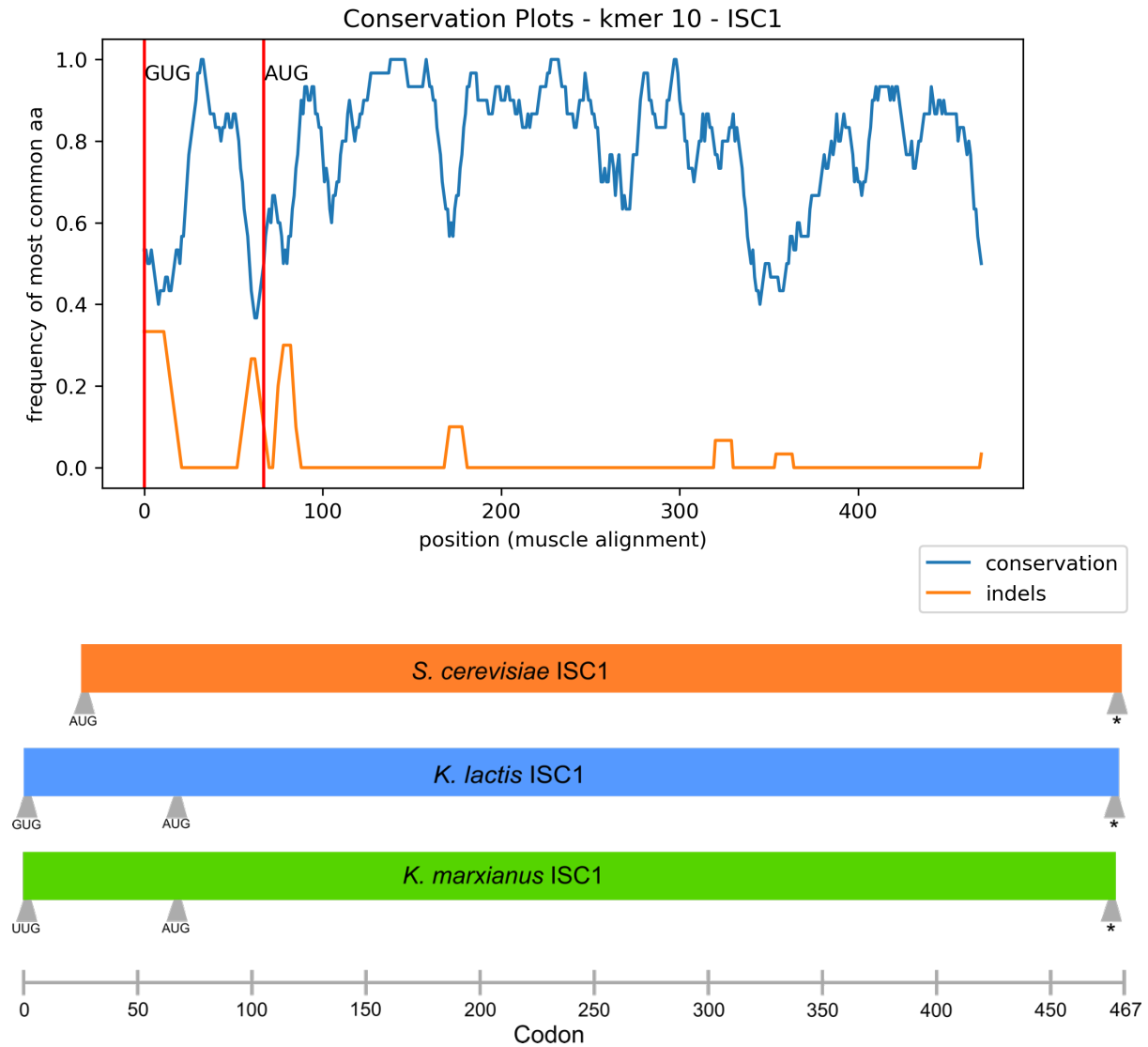


**Supplementary Figure 7.** *ADO1* Phobius prediction for extended protein (top) and truncated protein (bottom).

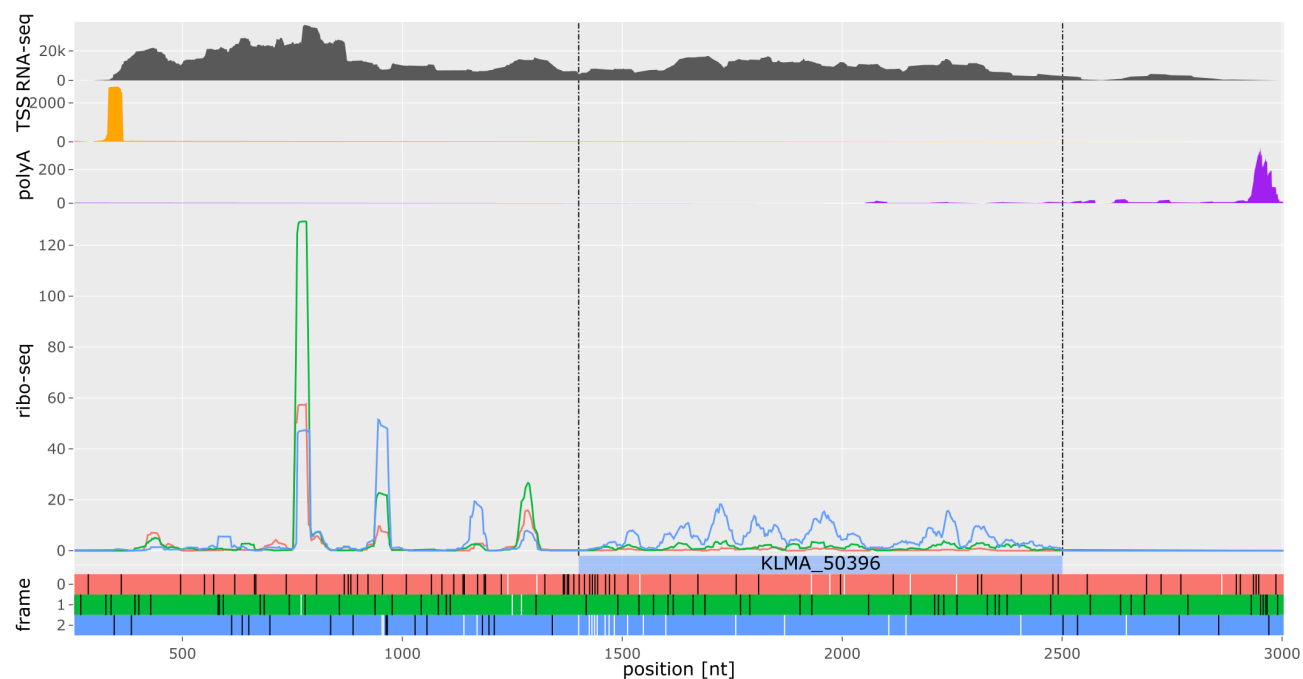


**Supplementary Figure 8.** Trips-Viz transcript plot of *S. cerevisiae* *ISC1* locus. Data represents ribosome profiling, with red line representing *ISC1* CDS in red frame. Note the absence of in-frame translation at the annotated CDS start, while clear translation coverage begins at a downstream in-frame AUG start codon. Data represents multiple ribosome profiling studies with *S. cerevisiae*.

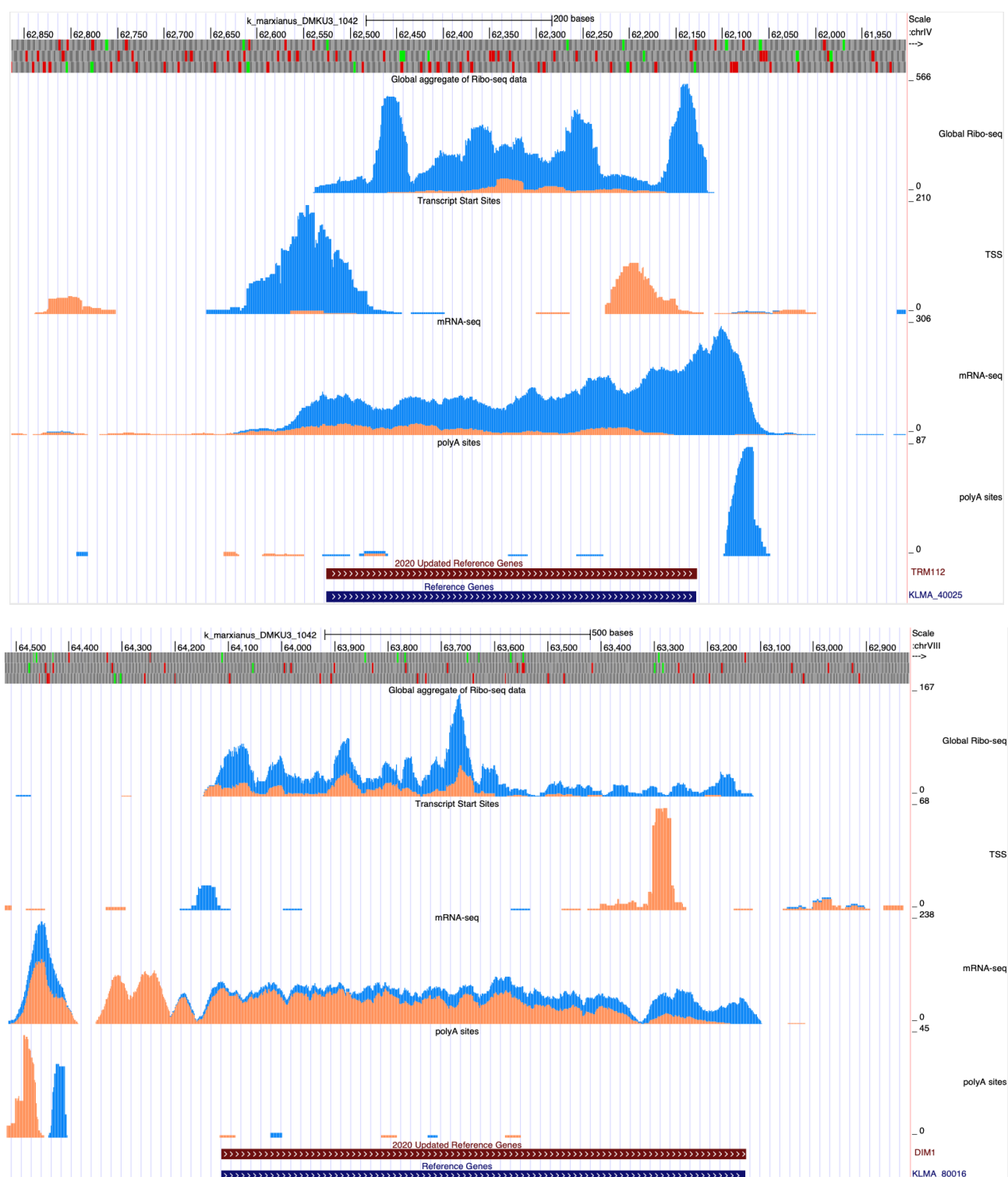




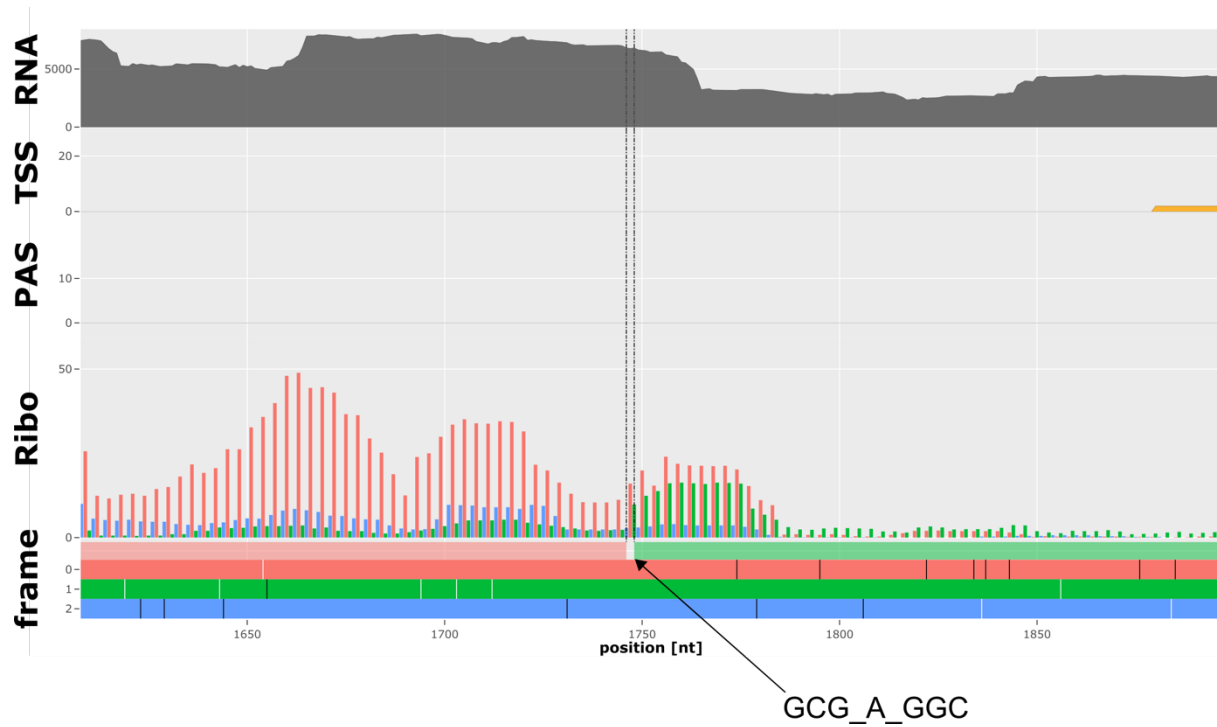
**Supplementary Figure 9.** Protein alignment and ORF architecture of *ISC1* homologs. **Top panel.** Alignment plot of *ISC1* homologous proteins from *K. marxianus*, *K. lactis* and *S. cerevisiae*. Relative positions of UUG and AUG start codons of *K. marxianus* are displayed in vertical red lines. Protein conservation between the red lines is present, indicating that translation from UUG in *K. marxianus* includes conserved amino acids sequences. A kmer of 10 amino acids was used to generate line plots for both conservation (blue) and indels (orange). **Bottom panel.** On the *K. marxianus* homolog (green), UUG represents exclusive non-AUG translation start site and AUG represents annotated start codon. GUG on *K. lactis* homolog (blue) is highlighted as a potential upstream in-frame start site and AUG represents annotated start codon. AUG on *S. cerevisiae* homolog (orange) represents true translation start site which was incorrectly annotated in strain S288c.



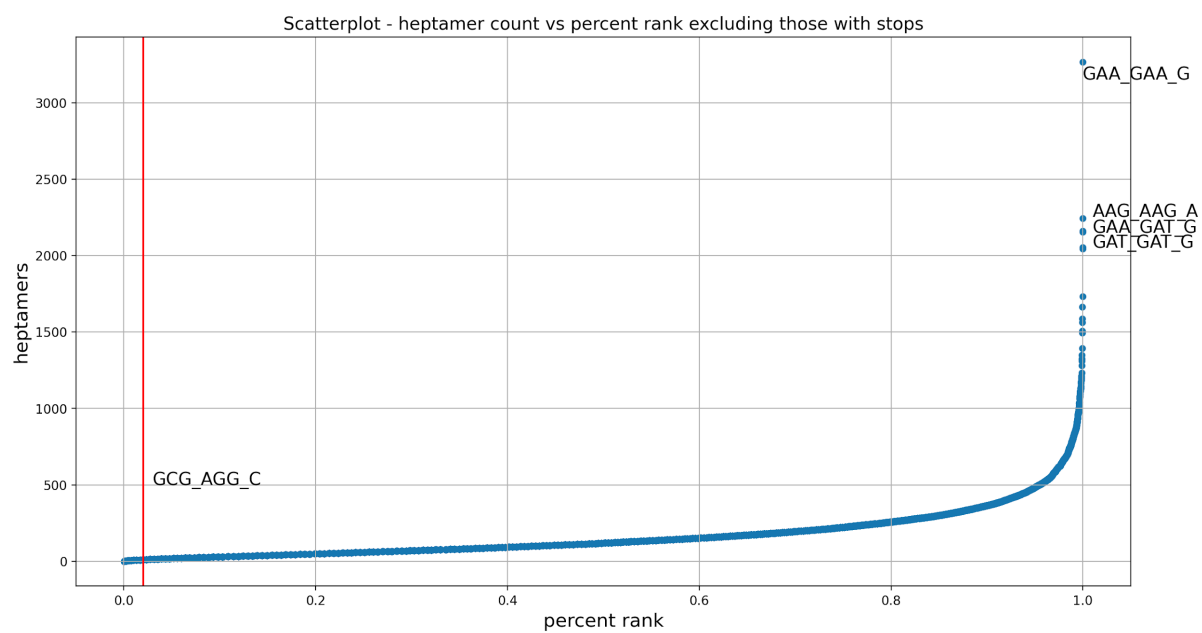
**Supplementary Figure 10.** Multiomics plot of the *GCN4* locus. The *GCN4* CDS is encoded by blue frame with multiple translated uORFs upstream. Note multiple translated uORFs in the 5' leader of the *GCN4* mRNA.



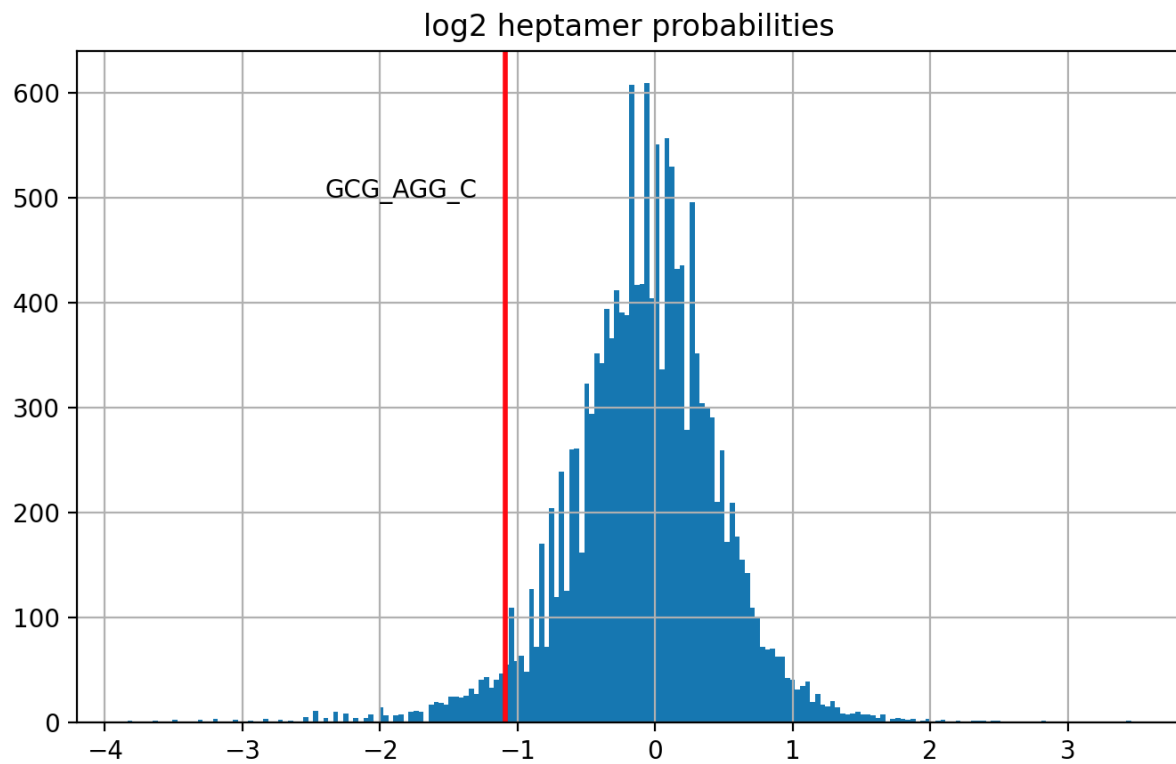
**Supplementary Figure 11.** GWIPS-Viz screenshot of *TRM112* locus (top) and *DIM1* (bottom). Blue reads represent the sense strand (annotated gene), orange reads represent antisense strand for both genes.



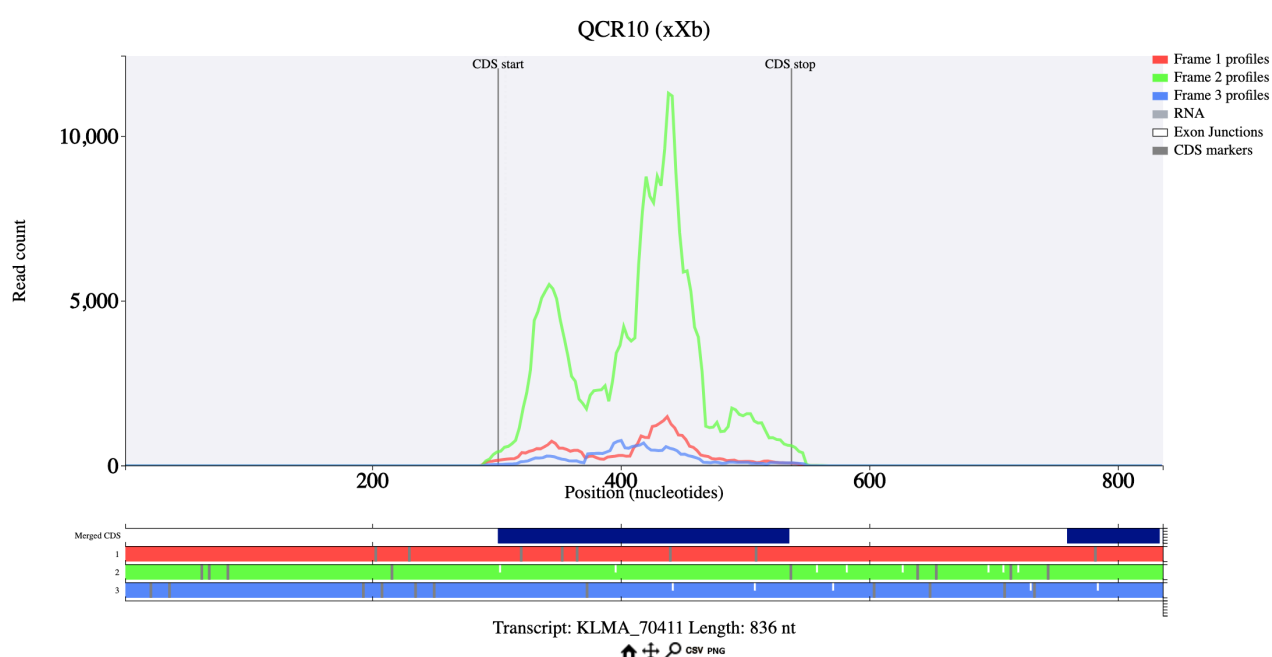
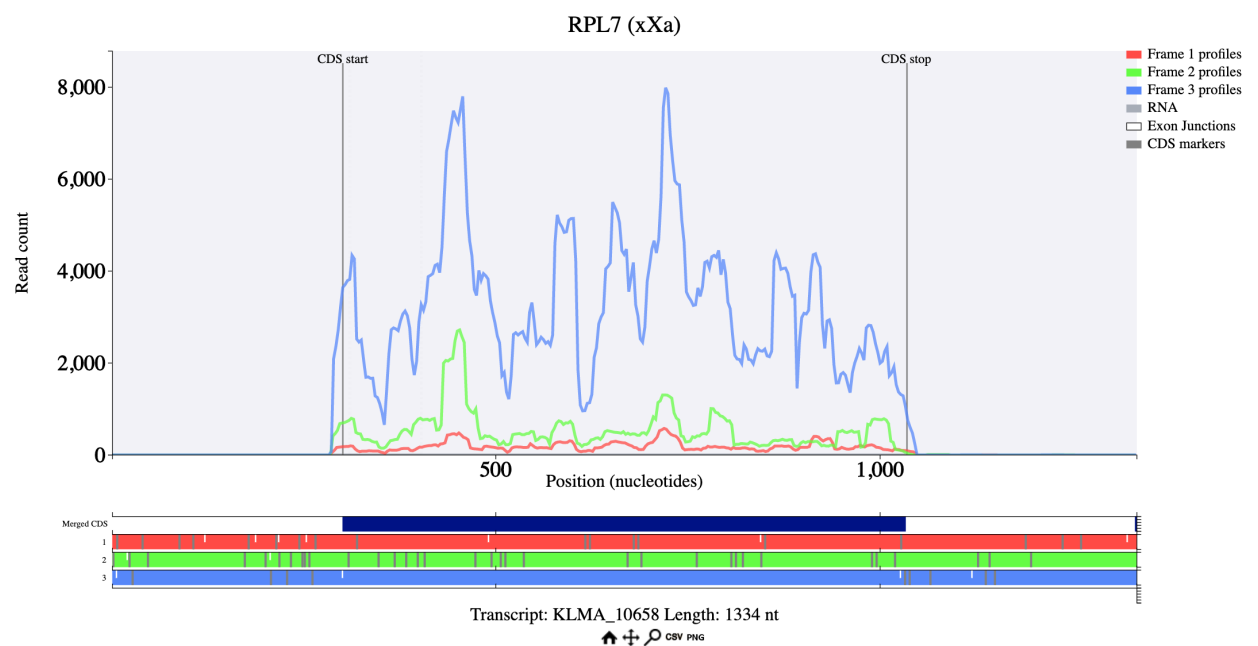
**Supplementary figure 12.** Multiomics plot of +1 frameshift locus. Note the increase in +1 frame translation (green columns) immediately downstream of the +1 frameshift site.



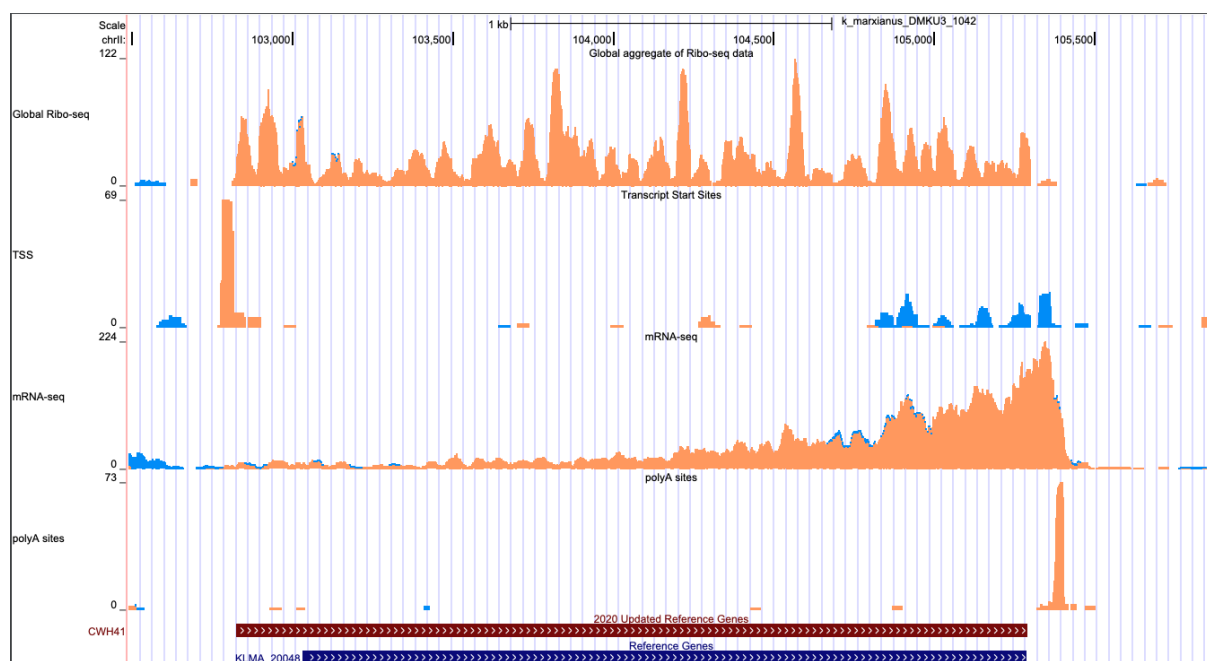
**Supplementary Figure 13.** Scatterplot displaying each CDS heptamer by percent rank. Annotated are the top three heptamers found in the CDS. The hybrid frameshift candidate (GCG\_AGG\_C) is marked with a vertical red line.



**Supplementary Figure 14.** Histogram displaying log2 heptamer probabilities of heptamers present in CDS regions. Frameshift candidate GCG\_AGG\_C is annotated with a vertical red line in the negative tail of the distribution.

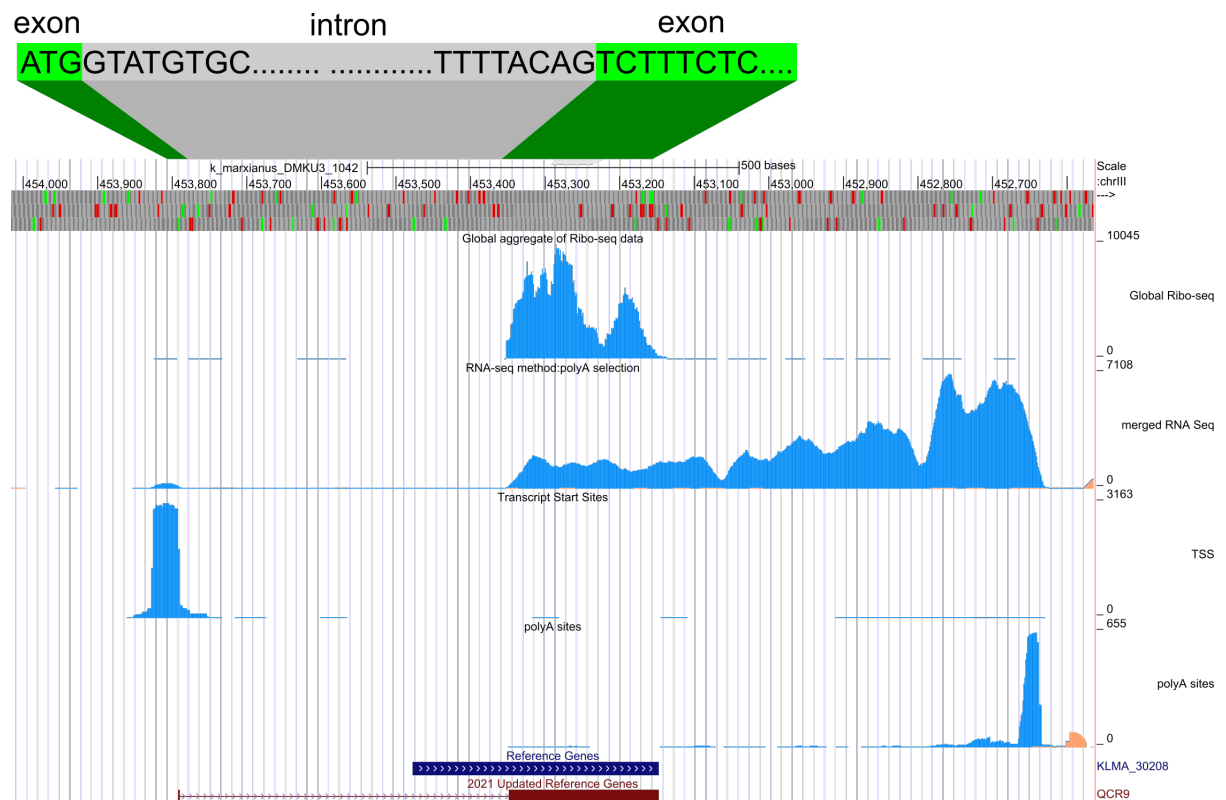


**Supplementary Figure 15.** Transcript plots of two discovered genes RPL7 (top) and QCR10 (bottom) generated from Trips-Viz. Blue/Green/Red horizontal bar corresponds to open reading frames. Vertical grey lines represent stop codons, white represent start codons.

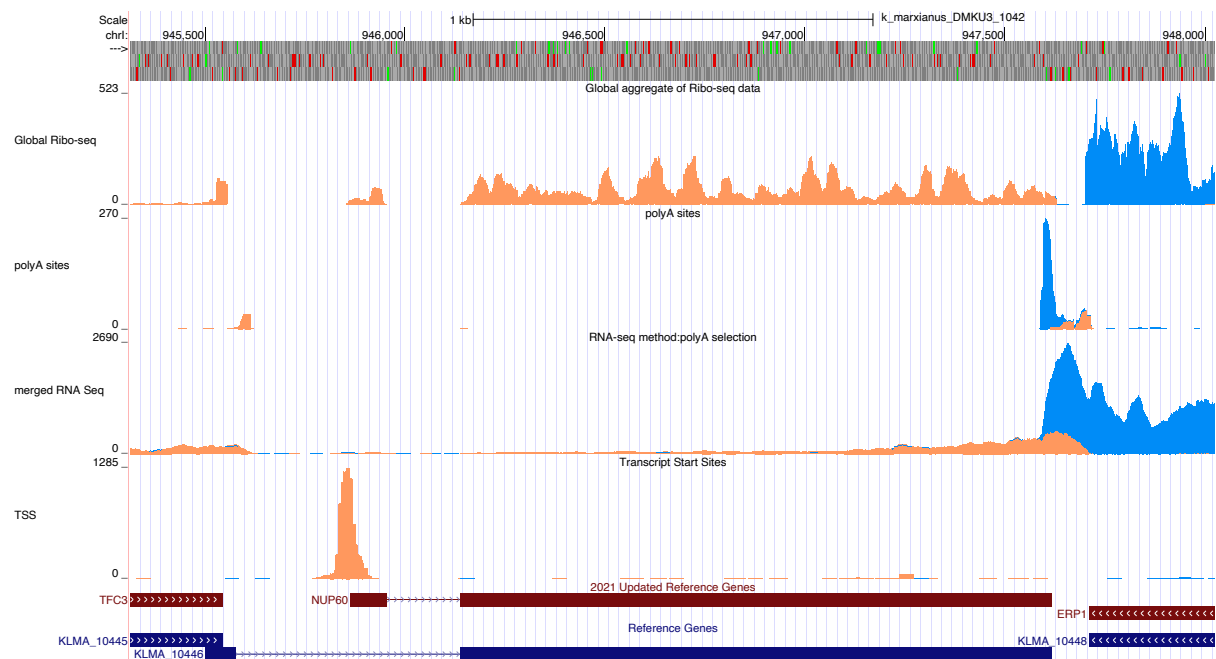


**Supplementary Figure 16.** Example of incorrect start codon assignment surrounding CWH41 locus. Note the original CDS annotation (blue bar) begins downstream of the translation start site, CDS start is corrected in updated annotation.

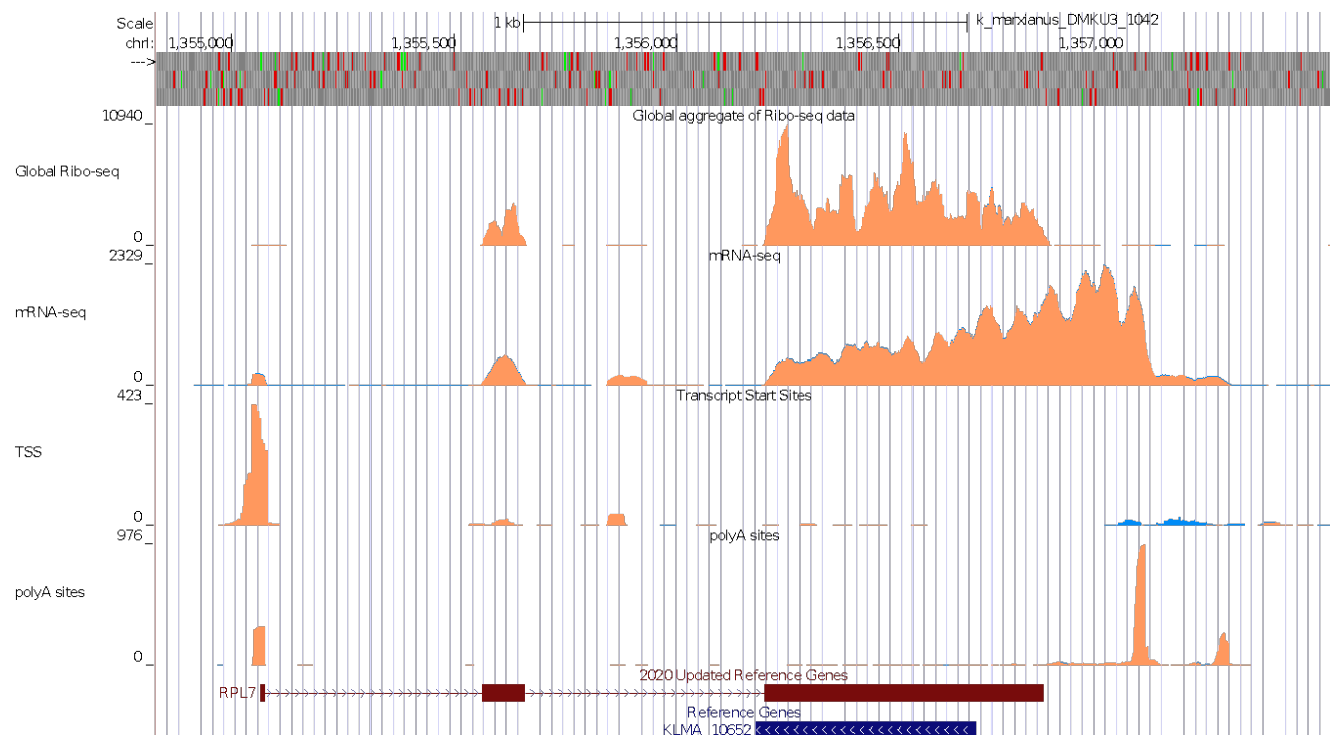




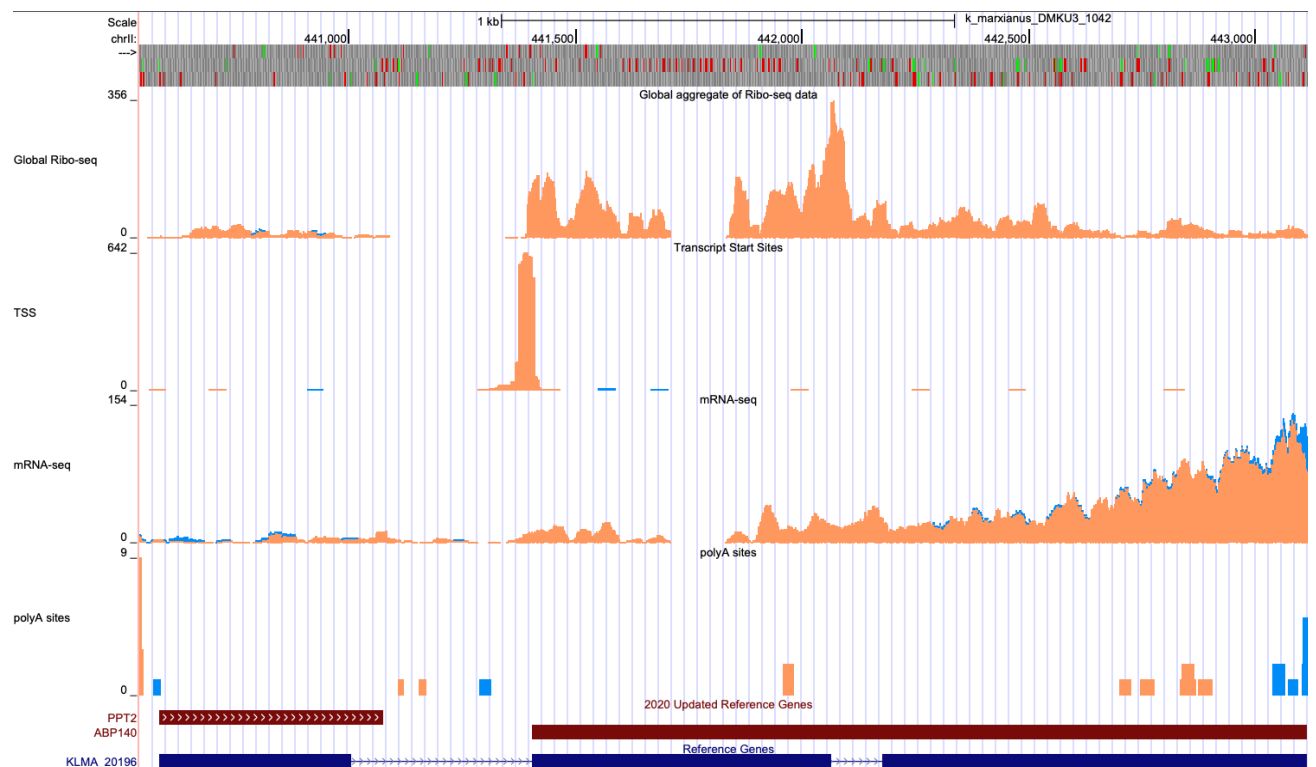
**Supplementary Figure 17.** Example of incorrect gene annotation. Genome browser screenshot surrounding KLMA\_30206 (*QCR9*). Original annotation corresponds to blue bar (“Reference Genes”) and red bar corresponds to corrected annotation including intron (“2020 Updated Reference Genes”). Sequence above browser represents the intron/exon junctions of *QCR9*.



**Supplementary Figure 18.** Example of incorrect intron/exon annotation in the NUP60 locus. Note the position of the first exon has been corrected in the new annotation. Previous annotation incorrectly annotated the first exon overlapping the *TFC3* gene CDS.



**Supplementary Figure 19.** Example of incorrect gene annotation (wrong strand). Genome browser screenshot surrounding KLMA\_10660 (RPL7). Original annotation corresponds to blue bar (“Reference Genes”) and red bar corresponds to corrected annotation including introns and exons (“2020 Updated Reference Genes”). Note the arrows showing which strand the gene annotation is orientated.



**Supplementary figure 20.** Example of incorrect gene annotation. Note the original CDS annotations created a single gene with exons encoding components of *PPT2* and the +1 frameshift gene *ABP140*. In updated annotation, *PPT2* has been assigned as an independent gene and *ABP140* has been given correct annotation for the 0-frame and +1 frame exon.

## **Chapter 4 - Heat-shock Induces a Rapid Increase in the Genes Involved in Cellular Respiration in the Yeast *Kluyveromyces marxianus***

### **Introduction**

In natural environments, yeasts respond to a wide-range of stresses such as oxidative, osmotic and heat stresses, that significantly alters their gene expression to allow an adaption to ensure growth and survival. Exposures to stresses also occurs in industrial settings, these include the heat stress associated with industrial-scale bioreactors, where large volumes of microbial activity generates large quantities of heat. Much of what we know about the response to heat stress derives from studies in the budding yeast *Saccharomyces cerevisiae* (reviewed in (Morano, Grant and Moye-Rowley 2012)). In *S. cerevisiae*, temperatures exceeding 37°C activate a transcriptional program known as the heat shock response (HSR) which results in the remodelling of gene expression. This HSR is activated via the heat shock transcription factor protein family (HSF), encoded by the genes *HSF1*, *MSN2* and *MSN4*. A major component of the HSR is the expression of a range of widely conserved heat shock proteins (HSPs) including Hsp26, which acts as a molecular chaperone for proteins preventing their denaturation and irreversible aggregation (Haslbeck *et al.* 1999). Other responses include increased concentrations of trehalose, which acts as a stabilizer of both proteins and membranes (Singer and Lindquist 1998), reorganisation of both cell membrane structure (Guyot *et al.* 2015) and cell wall (Hiromi and Hiroshi 2005), and the halting of cell growth via G1 arrest of the cell cycle (Rowley *et al.* 1993; Trotter *et al.* 2001).

Thermotolerance defines an organisms ability to grow and survive at elevated temperatures. *S. cerevisiae* is generally not defined as being thermotolerant as cells exposed to temperatures greater than 42°C are unable to cope for long periods. Studies have shown crucial enzymes such as RNA polymerase II (responsible for mRNA precursor synthesis (Sentenac 1985)) are largely inactive at temperatures exceeding >42°C (Noritaka *et al.* 2008). In addition, growth rates are reduced at temperatures ranging from 37-42°C. Many yeast species have evolved to be thermotolerant with the ability to grow at temperatures exceeding 37°C. Examples of thermotolerant yeasts include *Kluyveromyces marxianus*, *Pichia kudriavzevii* (also known as *Issatchenkia orientalis*) and *Candida tropicalis* (Talukder *et al.* 2016).

*Kluyveromyces marxianus* is commonly described as a thermotolerant yeast with strains reported to tolerate growth at temperatures exceeding 50°C. In contrast to the well-studied *S. cerevisiae*, this species also possesses a number of other defining traits such as the ability to consume a range of carbon sources including lactose, cellobiose, xylose and inulin. The ability to assimilate lactose explains the extensive use of this yeast with the production of traditional dairy products such as kefir and cheese (Gethins *et al.* 2016; Coloretto *et al.* 2017). This broad utilization of various sugars also makes this species attractive to food and biotechnology industries, as it can use a greater range of consumable carbon sources than more frequently used species such as *S. cerevisiae* (Fonseca *et al.* 2008; Morrissey *et al.* 2015; Varela *et al.* 2017; Rajkumar *et al.* 2019; Karim, Gerliani and Aider 2020; Rajkumar and Morrissey 2020; Leonel *et al.* 2021). In addition, thermotolerance provides the ability to ferment and grow at higher temperatures, leading to decreased cooling capacity requirements and the reduction in other microbial contaminants which are unable to grow at higher temperatures. To date, *K. marxianus* has also been reported as the fastest dividing eukaryotic microbe, with doubling times as low as 45 minutes allowing a much higher accumulation of biomass in a shorter time, again offering further opportunities to reduce costs in an industrial setting (Groeneveld, Stouthamer and Westerhoff 2009). These wide-range of competitive traits make this species attractive to food and biotechnology industries over commonly used species such as *S. cerevisiae* (Fonseca *et al.* 2008; Morrissey *et al.* 2015; Varela *et al.* 2017; Rajkumar *et al.* 2019; Karim, Gerliani and Aider 2020; Rajkumar and Morrissey 2020; Leonel *et al.* 2021). The natural benefits of this strain have been complemented by the establishment of tools to allow engineering of the genome, allowing it to be further optimized for industrial scale use, including tool kits for easy expression of multiple genes and kits utilizing CRISPR/Cas9 machinery (Cernak *et al.* 2018; Rajkumar *et al.* 2019).

Due to these preexisting industrial traits, there is a rationale to study how *K. marxianus* adapts to high temperatures in order to both understand and exploit these traits with a view towards developing strains better suited for industrial use. Many recent studies have focused on the use of RNA-sequencing to understand gene expression responses, examples include ethanol tolerance during adaptive laboratory evolution (Mo *et al.* 2019) and response to growth inhibitors derived from lignocellulosic substrates (Wang *et al.* 2018). There has also been an investigation into the novel genes specific to *K. marxianus* and how these are overrepresented

at high temperatures in an industrial setting (Doughty *et al.* 2020). As RNA-seq only measures the relative abundances of mRNAs, protein translation is ignored and the underlying assumption made is that a change in mRNA abundance reflects protein abundance. The use of ribosome profiling in *S. cerevisiae* has proven to be an extremely valuable tool in studying gene expression responses such as the HSR (Mühlhofer *et al.* 2019), oxidative stress (Blevins *et al.* 2019) and the yeast meiotic program (Brar *et al.* 2012). This technique isolates and sequences the location of translating ribosomes at codon resolution genome-wide (Ingolia *et al.* 2009), providing a method to quantify gene expression at a translational level rather than a transcriptional level. In addition, ribosome profiling has shown to be a more accurate measure of protein abundance than RNA-seq (Blevins *et al.* 2019).

We previously developed a ribosome profiling protocol along with a number of complimentary publicly available tools allowing ribosome profiling and RNA-seq data to be easily displayed (Fenton *et al.* 2022a). In addition, we developed an updated version of the *K. marxianus* genome annotation, resulting in the addition of over 170 novel genes (Fenton *et al.* 2022b) to the *K. marxianus* strain DMKU3-1042 genome (Lertwattanasakul *et al.* 2015). Here, using our updated annotation, a combination of ribosome profiling and RNA-seq was used to investigate how *K. marxianus* adapts to a rapid temperature shift from 30°C to 40°C, utilising timepoints from early as five minutes to one hour post heat-shock. This analysis revealed a response at 5 minutes post heat shock initiation which suggests cells upregulate cellular respiration in response to heat as described within.

## Results

### Ribosome Profiling Timecourse Reveals the Early Heat Shock Response.

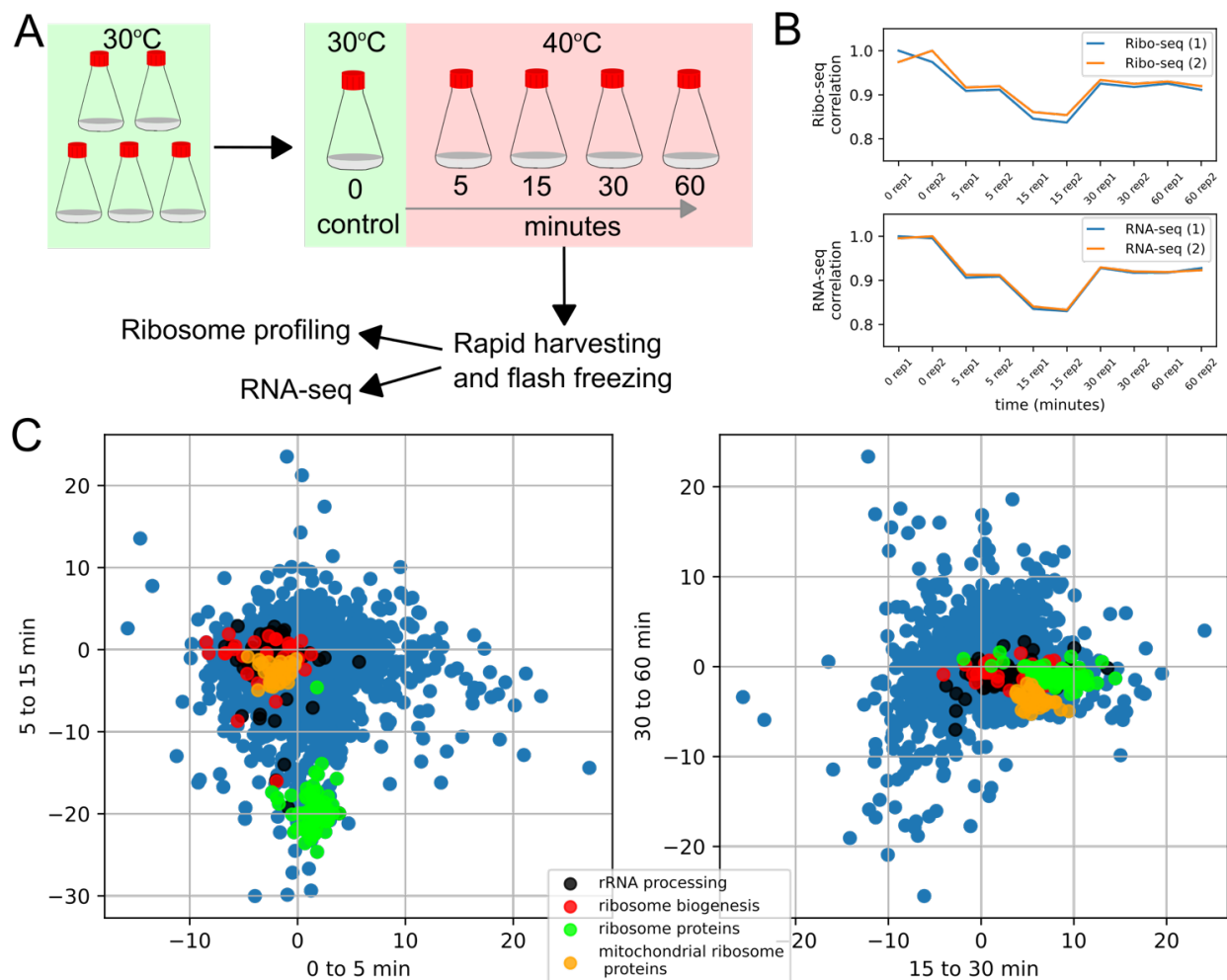
For our study, we decided to focus on relatively early timepoints to understand how *K. marxianus* responds to a rapid increase in temperature. Thus, we designed and carried out the following experiment whereby cultures growing at 30°C were transferred to a water bath at 40°C, samples were incubated and rapidly harvested at 5, 15, 30 and 60 minutes (Figure 1A). Both RNA-seq and Ribo-seq was carried out on each sample to determine gene expression changes during the timecourse. As Ribo-seq provides a more accurate account of gene expression changes, it was decided to use ribosome profiling data to determine gene expression changes throughout the timecourse.

Investigating gene expression correlations, both RNA-seq and Ribo-seq data suggested the greatest change in gene expression occurs at 15 minutes relative to the 30°C control (Figure 1B). In this study, we focused predominantly on the two earliest timepoints, 5 and 15 minutes after the initial 30-40°C heat shock. At the outset we targeted major gene groups which may give information on the cellular growth state during 5 and 15 minutes, particularly ribosome biogenesis, which is a major consumer of cellular energy during exponentially growing yeast, with ~60% of total transcription devoted to production of ribosomal RNA (Warner 1999). Interestingly, it was initially observed that genes involved in rRNA synthesis were being downregulated at 5 minutes, while ribosome proteins were downregulated at 15 minutes (Figure 1C), suggesting the regulation of rRNA synthesis and ribosome protein synthesis are asynchronous. We also observed large down-regulation of genes involved in nucleotide synthesis and DNA replication processes (see supplementary figure 1). Together, the down-regulation of rRNA, ribosome protein, nucleotide synthesis suggested an immediate large-scale response to increased temperatures at 5 minutes.

Next, Gene Ontology enrichment analysis (see methods) was performed to determine biological processes most effected by gene expression changes occurring at 5 minutes. As expected in a heat shock response, it was found that biological processes including protein folding, response to heat, cell redox homeostasis, trehalose biosynthesis and protein chaperones were significantly upregulated (Table 1). Downregulated gene ontology groups included ribosome biogenesis, nucleotide synthesis and DNA replication; suggesting a reduction of growth (Table 2). Surprisingly, many biological processes upregulated at 5



minutes included processes involved in cellular respiration (see table 1). These up-regulated biological processes at 5 minutes include glycolysis, TCA cycle, electron transport chain and ATP biosynthesis (Table1). Together these data suggested that cellular respiration was up-regulated at 5 minutes. In addition, it can be noted that the pentose phosphate pathway is also upregulated.



**Figure 1. Ribosome profiling and RNA-seq reveal changes in gene expression during heat shock.** A. Experimental design of heat shock timecourse. B. Spearman correlation of normalized ribosome counts (top) and RNA-seq counts (lower) between conditions and biological replicates using the 30°C data points as reference for correlation. C. Differential expression plots (stat plots) depicting gene expression changes throughout the timecourse measured as translation change. Left stat plot displays changes from 0 to 5 to 15 minutes. Right stat plot displays changes from 15 to 30 to 60 minutes. Stat values >3 or <-3 represent statistically significant gene expression changes (P-value <0.05)

	Biological Process	Genes	#genes	#DE genes	#total genes in BP	#genes	-log <sub>10</sub> p-value
Respiration	energy coupled proton transport, down electrochemical gradient	ATP4, ATP17, ATP16, atpH, ATP18, KLMX_50136, ATP3, ATP7, ATP2, VMA1, ATP1, KLMX_80023, KLMX_90008	13	710	19	5117	7.11
	ATP biosynthetic process	ATP4, ATP17, ATP16, atpH, ATP18, KLMX_50136, ATP3, ATP7, ATP2, VMA1, ATP1, KLMX_80023, KLMX_90008	13	710	19	5117	7.11
	glycolytic process	GPM3, ENO, PGK, RAG2, GPM1, TPI1, FBA1, RAG5, GAP3	9	710	13	5117	5.11
	mitochondrial electron transport, ubiquinol to cytochrome c	QCR7, QCR8, QCR9, QCR2, QCR6, QCR10	6	710	7	5117	4.36
	mitochondrial electron transport, cytochrome c to oxygen	COX4, COX6, COX9, COX8, COX5A, COX7	6	710	7	5117	4.36
	tricarboxylic acid cycle	LSC2, SDH1, LSC1, SDH3, IDP2, KGD2, MDH1, MLS1, FUM1, IDH1	10	710	19	5117	4.15
	proton transmembrane transport	VMA3, COX9, COX8, QCR2, VMA4, NHA1, QCR10, COX7, COX2, COX1	10	710	25	5117	2.95
Stress	protein folding	DNAJA1, STI1, MPD1, SSC1, DNAJB13, dnaJ, CPR5, SBA1, HSP26, APJ1, CPR3, AHA1, HSP82, CPR1, CPR6, HSP10, DNAJA1, MZM1, HCH1	19	710	52	5117	4.47
	response to heat	PIL1, KLMX_20435, dnaJ, APJ1, LSP1, SGT2, DNAJA1	7	710	10	5117	4.10
	stress granule disassembly	SKY1, HSP104, CUZ1	3	710	3	5117	2.57
	protein localization to eisosome filament	SUR7, PIL1, LSP1	3	710	3	5117	2.57
	trehalose biosynthetic process	TPS2, TSL1, TPS1	3	710	3	5117	2.57
Other	eisosome assembly	PIL1, NCE102, LSP1, SLM1	4	710	5	5117	2.79
	mitochondria-associated ubiquitin-dependent protein catabolic process	DNAJB13, UBR1, UBX2, CDC48, UBC4	5	710	8	5117	2.70
	D-xylose catabolic process	XYL1, YPR1, KLMX_80176	3	710	3	5117	2.57
	translational elongation	HYP2, EFB1, STM1, CAM1, RPP1B, TEF	6	710	12	5117	2.51
	protein peptidyl-prolyl isomerization	RRD2, CPR5, CPR3, CPR1, CPR6, FPR1	6	710	12	5117	2.51
	pentose-phosphate shunt	SOL3, GND1, TAL1, RPE1, SOL1	5	710	6	5117	3.57
	ubiquitin-dependent ERAD pathway	HRD1, DNAJA1, DSK2, UBR1, DFM1, LCL2, UBX2, UFD2, DNAJA1, SHP1	10	710	25	5117	2.95
	Ribophagy	VAC8, UBP3, BRE5, CDC48	4	710	5	5117	2.79

**Table 1.** Table of top twenty upregulated Biological Processes. Leftmost column represents sorting of Biological Processes into Respiration, Stress or Other.

	Biological Process	gene_name	# DE genes	# total DE genes	# total genes in BP	# genes	-log10 P-value
Ribosome/rRNA synthesis	rRNA processing	KRR1, BUD21, EFG1, CGR1, UTP13, DBP7, UTP15, UTP11, RAT1, UTP14, MPP10, IPI3, MRT4, NHP2, SSF1, MDN1, UTP10, GAR1, RPF1, XRN1, IMP4, RIX1, TSR2, PWP1, UTP22, MRD1, POP1, LRP1, RRP8, DBP10, RRP15, RRP9, NSA2, NOP9, SPB4, POL5, UTP23, UTP25, ESF1	39	656	68	5117	17.68
	maturation of SSU-rRNA from tricistronic rRNA transcript	UTP5, RRP36, UTP30, HAS1, NAN1, DCAF13, BFR2, SLX9, ECM16, RRP12, UTP9, UTP6, ENP2, UTP4, NOP7, FAF1, RIO2, DHR2, RPS20	19	656	23	5117	13.32
	ribosomal small subunit biogenesis	EFG1, TMA23, UTP13, UTP15, UTP14, MPP10, YAR1, NOB1, UTP10, NOP19, TSR2, SGD1, RRP14, RRP9, NOP9, LTV1, UTP23	17	656	20	5117	12.35
	ribosomal large subunit biogenesis	SDA1, TIF6, DBP7, NSA1, RAT1, REI1, NOP16, ALB1, RRP14, KLMX_70062, JIP5, RRP15, LOC1, PUF6	14	656	17	5117	9.87
	ribosomal large subunit assembly	RPF2, NOP53, IPI3, MRT4, IPI1, MDN1, RIX1, SQT1, YVH1, MAK21, SPB4, POL5, HPM1	13	656	15	5117	9.73
	ribosome biogenesis	RRP43, RCL1, EBP2, SNM1, NOG1, NOP58, NOP12, NOC3, RRB1, BMS1, MRM1, CSL4, RTC6, TSR1, NOP14, NOC4, RLP24, NOP2, RIO1	19	656	31	5117	9.56
	maturation of LSU-rRNA from tricistronic rRNA transcript	HAS1, NOP53, DBP6, RLP7, NOP8, IPI1, MAK5, HIT1, CIC1, ERB1, YTM1, NUG1, NOP7, NOP15	14	656	18	5117	9.27
	endonucleolytic cleavage in ITS1 to separate SSU-rRNA from 5.8S rRNA and LSU-rRNA from tricistronic rRNA transcript	ENP1, UTP7, PWP2, KRI1, UTP18, DIP2, FCF2, BUD23, SAS10, LCP5, LOC1	11	656	13	5117	8.06
	nuclear polyadenylation-dependent rRNA catabolic process	DIS3, RAT1, RRP40, RRP46, CSL4, LRP1, RRP4, RAI1, RRP6	9	656	12	5117	5.86
	maturation of 5.8S rRNA from tricistronic rRNA transcript	DIS3, DBP6, MAK5, CIC1, ERB1, YTM1, NOP19, UTP6, NOP7	9	656	13	5117	5.41
	assembly of large subunit precursor of preribosome	RPF2, NOG1, NIP7, DBP10, SPB4, RLP24	6	656	6	5117	5.36
	endonucleolytic cleavage in 5'-ETS of tricistronic rRNA transcript	UTP7, PWP2, UTP18, DIP2, FCF2, SAS10, LCP5, LOC1	8	656	11	5117	5.09
	rRNA metabolic process	RRP43, SNM1, CCM1, MRM1, NOP14	5	656	5	5117	4.47
	endonucleolytic cleavage to generate mature 5'-end of SSU-rRNA	UTP7, PWP2, UTP18, DIP2, FCF2, SAS10, LOC1	7	656	10	5117	4.33
Other	RNA processing	RRP43, RCL1, SNM1, MRPL15, MRPL3, MRM1, CSL4, PUS4, NOP2	9	656	14	5117	5.01
	amino acid transport	CAN1, LYP1, GAP1, TAT2, FSF1, SAM3, DIP5, PUT4, HIP1	9	656	15	5117	4.66
	Translation	NAM9, RSM22, MRPL22, MRP4, MRPL40, rplE, MRPS28, YML6, MRPL8, MRPL4, MRPS17, RTC6, RSM7, IMG2, MRPS5	15	656	39	5117	4.34
	nuclear polyadenylation-dependent tRNA catabolic process	DIS3, RRP40, SKI6, RRP46, CSL4, RRP4, RRP6	7	656	10	5117	4.33
	ncRNA 5'-end processing	UTP13, UTP14, MPP10, UTP10, NOP14, RRP9, NOP9, UTP23	8	656	8	5117	7.15
	'de novo' IMP biosynthetic process	ADE1, ADE4, ADE2, ADE13, ADE8, ADE5,7, ADE6	7	656	8	5117	5.41

**Table 2.** Table of top twenty downregulated Biological Processes. Leftmost column represents sorting of Biological Processes into Ribosome/rRNA synthesis or other.

### Cellular Respiration is Up-regulated at 5 minutes

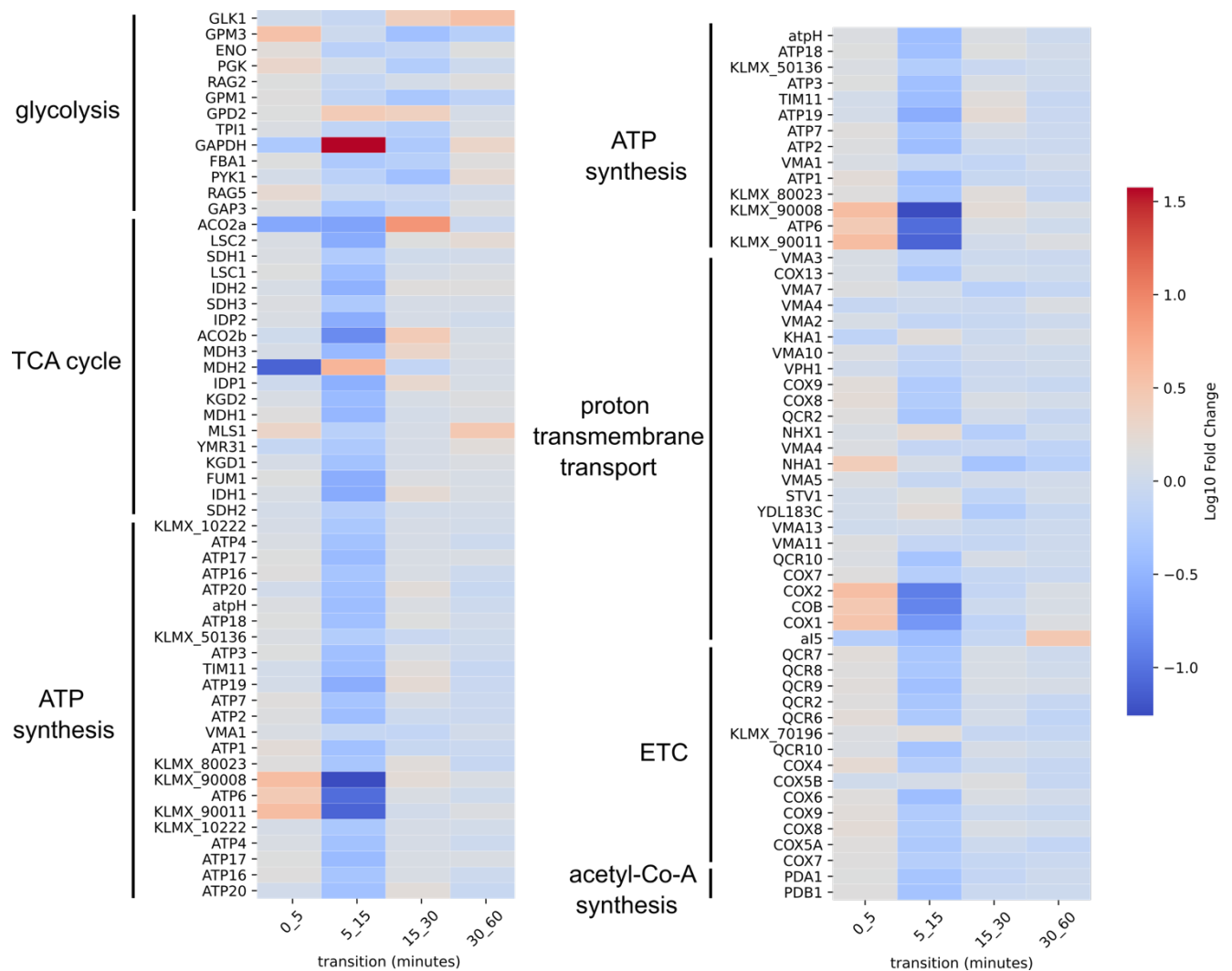
We focused on biological processes involved in cellular respiration. These included genes which function in cellular respiratory processes that includes glycolysis, TCA cycle, electron transport chain and ATP biosynthesis (Figure 4). It was found that the majority of the genes in these groups are significantly upregulated at 5 minutes. In parallel to GO-term analysis, we decided to investigate the most massively regulated genes, therefore prioritizing genes with large changes in regulation at 5 minutes post heat-shock. Genes are ranked from most significantly up and downregulated, as calculated by DESeq2 (see table 2). It is important to note that ~830 genes encoded in the *K. marxianus* genome are of unknown function (~20% of all genes), as such, these genes are expected to be overlooked when using GO term analysis and it can be argued that these should be analysed individually. Each massively regulated gene was independently analysed to determine potential function via homology to known proteins or functional domains. These include *FES1*, *CUZ1*, *ITR2* and *OPI10* as shown in Figure 3. *FES1* is necessary for the release of misfolded proteins from Hsp70 (Kabani, Beckerich and Brodsky 2002), Cuz1 binds to proteasomes and has roles in ubiquitin dependent protein degradation (Hanna *et al.* 2014), Itr2 encodes a myo-inositol transporter located at the cell membrane and vacuole (Nikawa, Tsukagoshi and Yamashita 1991; Nikawa, Hosaka and Yamashita 1993) while Opi10 is suggested to play a role in repression of genes involved in phospholipid biosynthesis (LC, RP and JM 2006).

An example of such a gene which is suggested to play an important role in this respiratory response is *TYE7*, translation of this gene increases ~7.8 fold ( $p < 0.01$ ). This gene is assigned to GO term groups too small to be significantly regulated. *TYE7* is a transcription factor responsible for the activation of glycolytic genes (Sato *et al.* 1999). This data would suggest *TYE7* is upregulated early in the heat shock response and may be responsible for the transcriptional activation of glycolytic genes at 5 minutes of our heat shock timecourse.

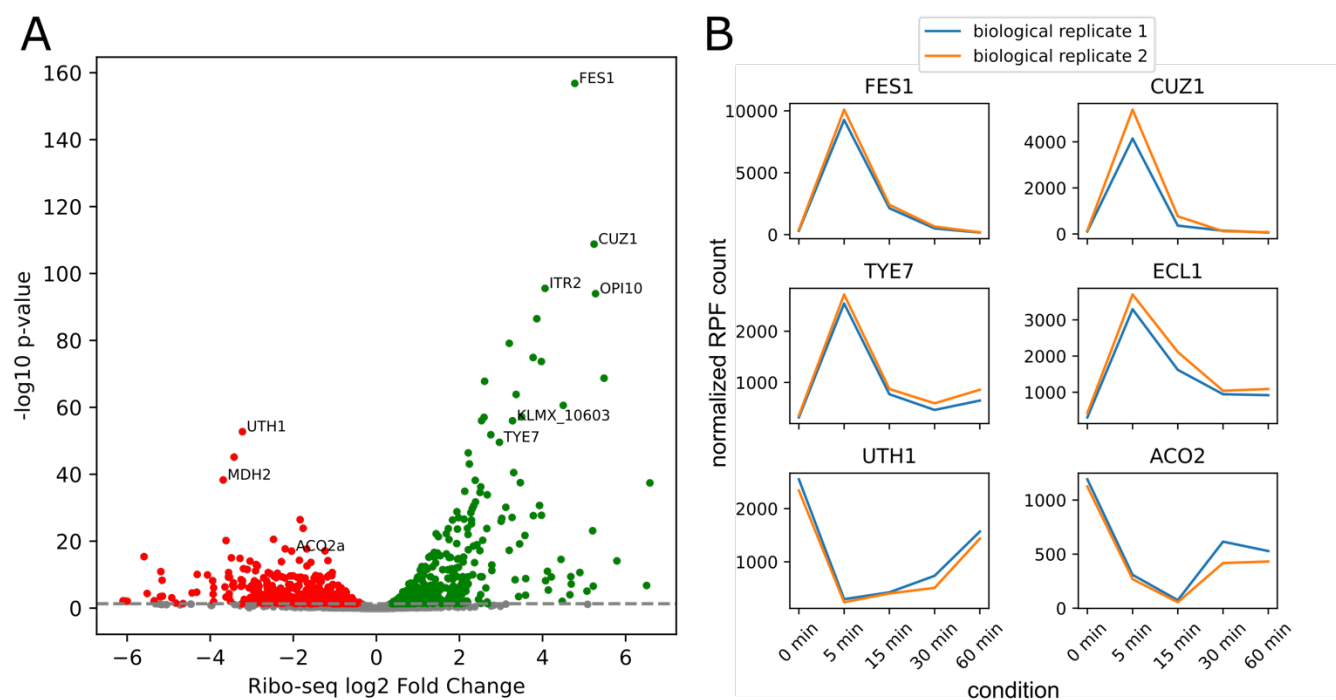
One gene that does not possess a gene name (such as *HSP26*), is only referred to as the gene ID/locus tag KLMX\_10603 and was upregulated at 5 minutes. Our GO software did not assign any GO term to this gene. Using blastp against the nr (non-redundant database), we determined this gene is a homolog of *S. cerevisiae* Ecl1. With 61% identity to the *Kluyveromyces lactis* and 16% identity to *S. cerevisiae* Ecl1 homologs. While the exact role of *ECL1* is unknown, many studies have elucidated its potential roles. Notably, one study

discovered respiratory activity and oxygen consumption of cells increased when the *Ecl1* was overexpressed (Azuma *et al.* 2012). In addition, this gene has been identified as a multicopy suppressor of temperature sensitive Hsf1 mutants of *S. cerevisiae*.

The *MDH* gene family encodes proteins that function as malate dehydrogenases, encoded between three isoforms *MDH1*, *MDH2* and *MDH3* (McAlister-Henn and Thompson 1987; Minard and McAlister-Henn 1991; Steffan and McAlister-Henn 1992). *MDH2* encodes a cytoplasmic malate dehydrogenase, that catalyses the interconversion of malate and oxaloacetate and is involved in gluconeogenesis. In these data, it was observed this gene becomes massively downregulated at 5 minutes. Moreover, the mitochondrial isoform is significantly upregulated, while there is no significant change in the peroxisomal *MDH3* isoform (Steffan and McAlister-Henn 1992). The differential expression of these MDH enzymes suggests gluconeogenesis is downregulated at 5 minutes while the TCA-cycle-dependent mitochondrial isoform is upregulated, allowing increased conversion of malate to oxalacetate, consistent with the global upregulation of aerobic respiration. Alongside *MDH2*, it was observed *ACO2* was significantly downregulated. *ACO1* and *ACO2* encode key components of the TCA cycle. The fermenting yeast *S. cerevisiae* has two encoded aconitase genes, *ACO1* that is essential for the citric acid cycle, and *ACO2* that has been shown to be specifically and exclusively contributing to lysine biosynthesis (Fazius *et al.* 2012). *ACO2* catalyzes the reversible dehydration of (R)-homocitrate to cis-homoaconitate, a step in the alpha-aminoadipate pathway for lysine biosynthesis. This specific downregulation of *ACO2* at 5 minutes would reduce exit flux of TCA cycle intermediates for lysine biosynthesis, allowing increased intermediates to complete the TCA cycle contributing to a potential increase in respiration. *UTH1* may be defined as the most significantly down-regulated gene at 5 minutes, which suggests reducing translation of this gene is important. This protein has been shown to locate to the mitochondria and has been suggested to play a role in mitophagy (Kissova *et al.* 2004), the process of autophagic degradation of mitochondria.



**Figure 2. Gene expression changes (log10 fold changes)** Biological processes directly involved in cellular respiration are upregulated at 5 minutes. Each heatmap represents gene expression changes through each progressive timepoint of the timecourse.

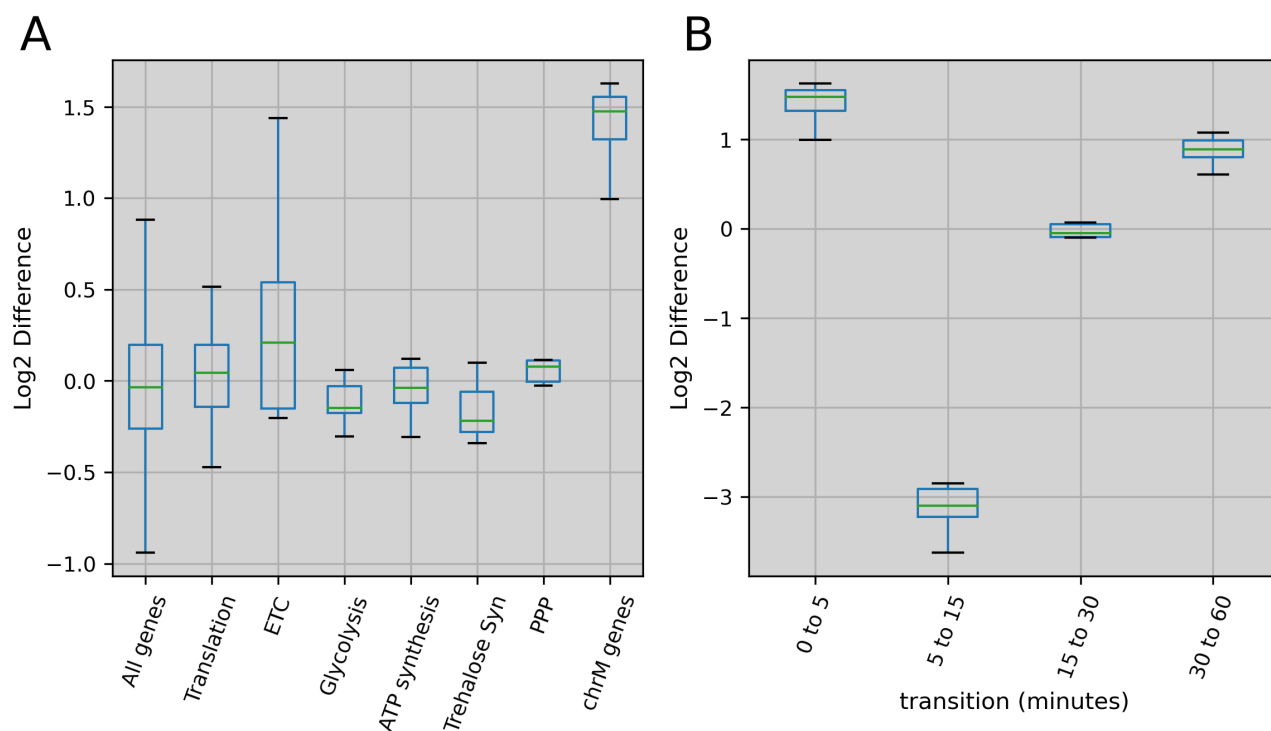


**Figure 3.** Investigation of massively regulated genes. A. Volcano plot of gene expression changes from 30°C to 40°C 5 minutes. Red dots represent statistically significant ( $p < 0.05$ ) genes which are downregulated at 5 minutes while green represents upregulated. Gene names are labelled to genes of interest. B. Normalized ribosome count values for each gene of interest plotted through each time-point.

### **Translation Efficiency Changes Suggest Mitochondrial mRNAs are Translationally Regulated.**

As both Ribo-seq and RNA-seq were carried out in parallel, it is possible to determine whether the change in gene expression is a result of transcription, translation or both processes. Thus, a change in mRNA which reflects changes in translation would represent a gene transcriptionally regulated and a gene with little change in mRNA abundance and a large change in translation would represent a gene with potential translational regulation. The translation efficiency (TE) was calculated for groups of genes (including GO groups) during the timecourse. By studying differences in mRNA abundance from RPF abundance, it was observed that genes involved in glycolysis, ATP synthesis and the electron transport chain are transcriptionally regulated (Figure 4A). However, studying the TE changes from mitochondrially encoded genes (chrM), a huge increase in TE from 30°C to 40°C 5 minutes was observed. These chrM genes include genes directly involved in cellular respiration (*COX1*, *COX2*, *COX3*, *COB* and *ATP6* and *ATP9*). Moreover, the ATP synthase complex is composed of proteins encoded in both nuclear and mitochondrial DNA, these data would suggest nuclear encoded ATP synthase proteins are transcriptionally upregulated while mitochondrial encoded ATP synthase proteins are translationally upregulated.

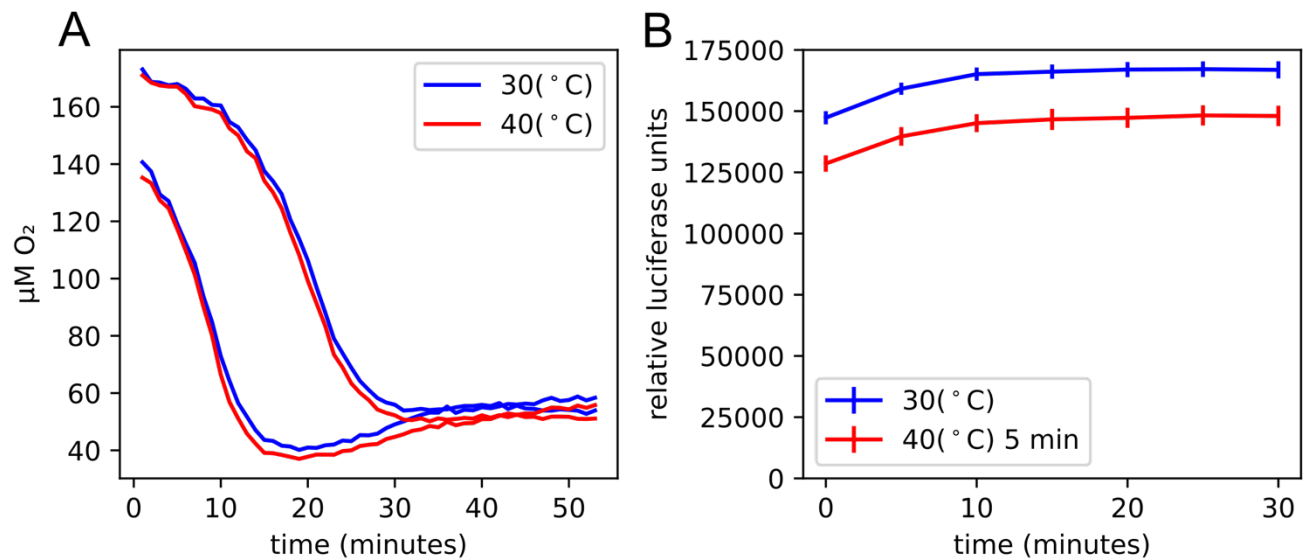




**Figure 4.** Translation efficiency changes during heat shock. A. Log2 difference in TE during transition from 30°C to 40°C 5 minutes. Note the increase in relative translation efficiency of mitochondria encoded genes. B. Translation efficiency changes of chrM genes during different transitions in the heat shock timecourse.

### Oxygen consumption and ATP assays

Taking advantage of a previously described oxygen consumption rate assay which was previously been used to study oxygen consumption in the fission yeast *Schizosaccharomyces pombe* (O’Riordan *et al.* 2000) (see methods), the oxygen consumption of both 5 minute (heat shocked) and 30°C cells (Figure 5A) was measured. While a significant increase in oxygen consumption could not be identified, we note a minor increase and thus rule out a decrease in oxygen consumption in heat shocked samples. Next, intracellular ATP concentrations using a luciferase-based reporter assay (see methods) was tested, to determine if ATP concentration changed in the 5 minute heat shocked samples relative to the 30°C control. An approximately 13% decrease in cellular ATP concentrations in the heat shocked samples was observed (Figure 5B). This assay suggests two potential reasons to decreased ATP levels, an increase in cellular ATP consumption or a decrease in ATP production. An increase in respiration may be a response to decreased ATP concentrations due to increased cellular energy demand.



**Figure 5.** Oxygen consumption rate and ATP assays. A. Oxygen consumption rate assay (OCR) for 30°C samples and 40°C 5 minute heat shocked samples. On this graph, two separate dilutions (1/4 and 1/8) are shown with the 1/4 dilution showing the fastest decrease in oxygen levels. Time (minutes) represents assay readings. B. ATP assay suggests lower ATP levels at and 40°C 5 minutes timepoint. Note the decrease in ATP concentration of 40°C 5 minute heat shocked samples. Time represents minutes post assay readings were started.

## Discussion

Studying differential gene expression has been fundamental to gaining an understanding of how an organism adapts to various stresses. Heat stress is of particular interest in *Kluyveromyces marxianus* as it is considered thermotolerant, with its capability of growing comfortably above 37°C. In this study, the transcriptional and translational response to heat-shock in the thermotolerant budding yeast *Kluyveromyces marxianus* was investigated. Many studies have examined the transcriptional response to increased temperatures after adaptation, ignoring the initial changes which allow long-term adaptation (Lertwattanasakul *et al.* 2015; Doughty *et al.* 2020). This study reports a significant increase in the relative translation of genes involved in aerobic respiration at 5 minutes. This response is supported by both general GO analysis and also further investigation into genes which may play important roles in cellular respiration (such as *TYE7*). Remarkably, we observe the upregulation of specific respiratory complexes such as the ATP synthase complex, which is comprised of multiple proteins, by two distinct mechanisms. Nuclear encoded genes involved in respiration appear transcriptionally upregulated, while mitochondrially encoded genes appear translationally upregulated (more ribosomes per mitochondrial mRNA), suggestive of mitochondrial-nuclear crosstalk. To support upregulation in cellular respiration, oxygen levels after 5 minutes of heat shock were measured. While this assay failed to produce significant evidence regarding an increase in oxygen, we note this assay takes approximately 15 minutes to prepare before readings commence and it's possible that any oxygen increase in missed. In addition to oxygen measurements, intracellular ATP concentrations showed a significant decrease in ATP levels after 5 minutes of heat shock. Decreased ATP levels are suggestive of a decrease or increase in ATP production rates. While further work will be required to support upregulation of respiration, we hypothesize that cells require a burst in energy to meet energy demands while adapting to heat stress.

## **Methods**

### **Ribosome Profiling and RNA-seq**

We used a modified version of the Ingolia ribosome profiling protocol (McGlinchy and Ingolia 2017; Fenton *et al.* 2022a). In brief, cells (CBS6556) were grown overnight in 5mL minimal media (Verduyn *et al.* 1992). 150 mL minimal media in 500 mL Erlenmeyer flasks were inoculated to an OD<sub>600</sub> 0.06. Cultures were incubated at 30°C with shaking (180 rpm) until cells reached ~0.6 OD<sub>600</sub>, before harvesting or transfer to a water bath at 40°C with shaking. Cultures were harvested at specific timepoints via rapid filtration and liquid nitrogen cooling. Cells were pulverized cryogenically with a mixer mill. Lysates were clarified and placed on a 10-50% sucrose gradient and monosomes were isolated using a Brandel UV fractionator. RNA was isolated with Trizol and ribosome footprints isolated on a 15% PAGE-urea gel. Prepared cDNA libraries were sequenced on Illumina HiSeq 3000 (GC3F, University of Oregon, Eugene, OR, USA). Adapters were removed using cutadapt (Martin 2011). Ribosomal RNA was removed using bowtie (Langmead 2010). non-rRNA aligned reads were aligned to the genome. Number of ribosomes per gene was performed with HTSEQ count (Anders, Pyl and Huber 2015).

### **Differential Gene Expression**

All differential gene expression was performed using DESeq2 (Love, Huber and Anders 2014).

### **GO Term Enrichment**

Gene ontology annotations were provided from Panzzer2 (Toronen, Medlar and Holm 2018). Fishers exact test was used to determine enrichment significance. GO term enrichment was performed on each condition using a stat threshold of +3 or -3 for up-regulated and down-regulated genes, respectively.

### **Oxygen Consumption Assay**

Cultures were prepared in the same method as ribosome profiling. After heat shock, both samples were allowed return to 30°C. A sample of culture was flash frozen for A600 and protein level analysis. Serial dilutions of culture were loaded onto a 96 well plate and sealed with mineral oil. Media without cells was used a negative control. A600 readings were measured in triplicate with a Tecan plate reader to normalize cell number differences between 30°C and 40°C cultures. In addition, Bradford Assay was used to determine protein level changes.

### ATP assay

ATP concentration was measured using the Promega Kinase-Glo assay using a 96-well white plate. Each well contained a mix of 50  $\mu$ L culture and 50  $\mu$ L Kinase-Glo solution and was carried out in triplicate. Luciferase activity was measured on a Turning Instruments luminometer with readings taken every 5 minutes for 30 minutes.

## References

- Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015;**31**:166–9.
- Azuma K, Ohtsuka H, Murakami H *et al.* Extension of chronological lifespan by ScEcl1 depends on mitochondria in *Saccharomyces cerevisiae*. *Biosci Biotechnol Biochem* 2012;**76**:1938–42.
- Blevins WR, Tavella T, Moro SG *et al.* Extensive post-transcriptional buffering of gene expression in the response to severe oxidative stress in baker's yeast. *Sci Rep* 2019;**9**:11005.
- Brar GA, Yassour M, Friedman N *et al.* High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* 2012;**335**:552–7.
- Cernak P, Estrela R, Poddar S *et al.* Engineering *Kluyveromyces marxianus* as a Robust Synthetic Biology Platform Host. *MBio* 2018;**9**, DOI: 10.1128/mBio.01410-18.
- Coloretti F, Chiavari C, Luise D *et al.* Detection and identification of yeasts in natural whey starter for Parmigiano Reggiano cheese-making. *Int Dairy J* 2017;**66**:13–7.
- Doughty TW, Domenzain I, Millan-Oropeza A *et al.* Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts. *Nat Commun* 2020;**11**:2144.
- Doyle SM, Wickner S. Hsp104 and ClpB: protein disaggregating machines. *Trends Biochem Sci* 2009;**34**:40–8.
- Fazius F, Shelest E, Gebhardt P *et al.* The fungal alpha-aminoadipate pathway for lysine biosynthesis requires two enzymes of the aconitase family for the isomerization of homocitrate to homoisocitrate. *Mol Microbiol* 2012;**86**:1508–30.
- Fenton DA, Kiniry SJ, Yordanova MM *et al.* Development of a Ribosome Profiling Protocol to Study Translation in the yeast *Kluyveromyces marxianus*. *bioRxiv* 2022a:2022.02.06.478964.
- Fenton DA, Świrski M, O'Connor PBF *et al.* Integrated data-driven reannotation of the <em>Kluyveromyces marxianus</em> genome reveals an expanded protein coding repertoire. *bioRxiv* 2022b:2022.03.25.485750.
- Fonseca GG, Heinzle E, Wittmann C *et al.* The yeast *Kluyveromyces marxianus* and its biotechnological potential. *Appl Microbiol Biotechnol* 2008;**79**:339–54.
- Gethins L, Rea MC, Stanton C *et al.* Acquisition of the yeast *Kluyveromyces marxianus* from unpasteurised milk by a kefir grain enhances kefir quality. *FEMS Microbiol Lett* 2016;**363**, DOI: 10.1093/femsle/fnw165.

- Groeneveld P, Stouthamer AH, Westerhoff H V. Super life--how and why “cell selection” leads to the fastest-growing eukaryote. *FEBS J* 2009;**276**:254–70.
- Guyot S, Gervais P, Young M *et al.* Surviving the heat: heterogeneity of response in *Saccharomyces cerevisiae* provides insight into thermal damage to the membrane. *Environ Microbiol* 2015;**17**:2982–92.
- Hanna J, Waterman D, Isasa M *et al.* Cuz1/Ynl155w, a zinc-dependent ubiquitin-binding protein, protects cells from metalloid-induced proteotoxicity. *J Biol Chem* 2014;**289**:1876–85.
- Haslbeck M, Walke S, Stromer T *et al.* Hsp26: a temperature-regulated chaperone. *EMBO J* 1999;**18**:6744–51.
- Hiromi I, Hiroshi S. *Saccharomyces cerevisiae* Heat Shock Transcription Factor Regulates Cell Wall Remodeling in Response to Heat Shock. *Eukaryot Cell* 2005;**4**:1050–6.
- Ingolia NT, Ghaemmaghami S, Newman JRS *et al.* Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 2009;**324**:218–23.
- Kabani M, Beckerich J-M, Brodsky JL. Nucleotide exchange factor for the yeast Hsp70 molecular chaperone Ssa1p. *Mol Cell Biol* 2002;**22**:4677–89.
- Karim A, Gerliani N, Aïder M. *Kluyveromyces marxianus*: An emerging yeast cell factory for applications in food and biotechnology. *Int J Food Microbiol* 2020;**333**:108818.
- Kissova I, Deffieu M, Manon S *et al.* Uth1p is involved in the autophagic degradation of mitochondria. *J Biol Chem* 2004;**279**:39068–74.
- Langmead B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinforma* 2010;**Chapter 11**:Unit-11.7.
- LC H, RP B, JM L. Genomic analysis of the Opi- phenotype. *Genetics* 2006;**173**:621–34.
- Leonel L V, Arruda P V, Chandel AK *et al.* *Kluyveromyces marxianus*: a potential biocatalyst of renewable chemicals and lignocellulosic ethanol production. *Crit Rev Biotechnol* 2021;**41**:1131–52.
- Lertwattanasakul N, Kosaka T, Hosoyama A *et al.* Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. *Biotechnol Biofuels* 2015;**8**:47.
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
- Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads.

- EMBnet.journal* 2011;**17**:10–2.
- McAlister-Henn L, Thompson LM. Isolation and expression of the gene encoding yeast mitochondrial malate dehydrogenase. *J Bacteriol* 1987;**169**:5157–66.
- McGlinchy NJ, Ingolia NT. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* 2017;**126**:112–29.
- Minard KI, McAlister-Henn L. Isolation, nucleotide sequence analysis, and disruption of the MDH2 gene from *Saccharomyces cerevisiae*: evidence for three isozymes of yeast malate dehydrogenase. *Mol Cell Biol* 1991;**11**:370–80.
- Mo W, Wang M, Zhan R *et al.* *Kluyveromyces marxianus* developing ethanol tolerance during adaptive evolution with significant improvements of multiple pathways. *Biotechnol Biofuels* 2019;**12**:63.
- Morano KA, Grant CM, Moye-Rowley WS. The Response to Heat Shock and Oxidative Stress in *Saccharomyces cerevisiae*. *Genetics* 2012;**190**:1157 LP – 1195.
- Morrissey JP, Etschmann MMW, Schrader J *et al.* Cell factory applications of the yeast *Kluyveromyces marxianus* for the biotechnological production of natural flavour and fragrance molecules. *Yeast* 2015;**32**:3–16.
- Mühlhofer M, Berchtold E, Stratil CG *et al.* The Heat Shock Response in Yeast Maintains Protein Homeostasis by Chaperoning and Replenishing Proteins. *Cell Rep* 2019;**29**:4593–4607.e8.
- Nikawa J, Hosaka K, Yamashita S. Differential regulation of two myo-inositol transporter genes of *Saccharomyces cerevisiae*. *Mol Microbiol* 1993;**10**:955–61.
- Nikawa J, Tsukagoshi Y, Yamashita S. Isolation and characterization of two distinct myo-inositol transporter genes of *Saccharomyces cerevisiae*. *J Biol Chem* 1991;**266**:11184–91.
- Noritaka Y, Yuka M, Aya I *et al.* Regulation of Thermotolerance by Stress-Induced Transcription Factors in *Saccharomyces cerevisiae*. *Eukaryot Cell* 2008;**7**:783–90.
- O’Riordan TC, Buckley D, Ogurtsov V *et al.* A Cell Viability Assay Based on Monitoring Respiration by Optical Oxygen Sensing. *Anal Biochem* 2000;**278**:221–7.
- Rajkumar AS, Morrissey JP. Rational engineering of *Kluyveromyces marxianus* to create a chassis for the production of aromatic products. *Microb Cell Fact* 2020;**19**:207.
- Rajkumar AS, Varela JA, Juergens H *et al.* Biological Parts for *Kluyveromyces marxianus* Synthetic Biology. *Front Bioeng Biotechnol* 2019;**7**:97.



- Rowley A, Johnston GC, Butler B *et al.* Heat shock-mediated cell cycle blockage and G1 cyclin expression in the yeast *Saccharomyces cerevisiae*. *Mol Cell Biol* 1993;**13**:1034–41.
- Sato T, Lopez MC, Sugioka S *et al.* The E-box DNA binding protein Sgc1p suppresses the *gcr2* mutation, which is involved in transcriptional activation of glycolytic genes in *Saccharomyces cerevisiae*. *FEBS Lett* 1999;**463**:307–11.
- Sentenac A. Eukaryotic RNA polymerases. *CRC Crit Rev Biochem* 1985;**18**:31–90.
- Singer MA, Lindquist S. Thermotolerance in *Saccharomyces cerevisiae*: the Yin and Yang of trehalose. *Trends Biotechnol* 1998;**16**:460–8.
- Sorger PK. Yeast heat shock factor contains separable transient and sustained response transcriptional activators. *Cell* 1990;**62**:793–805.
- Steffan JS, McAlister-Henn L. Isolation and characterization of the yeast gene encoding the MDH3 isozyme of malate dehydrogenase. *J Biol Chem* 1992;**267**:24708–15.
- Talukder AA, Easmin F, Mahmud SA *et al.* Thermotolerant yeasts capable of producing bioethanol: isolation from natural fermented sources, identification and characterization. *Biotechnol Biotechnol Equip* 2016;**30**:1106–14.
- Toronen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res* 2018;**46**:W84–8.
- Trotter EW, Berenfeld L, Krause SA *et al.* Protein misfolding and temperature up-shift cause G<sub>1</sub> arrest via a common mechanism dependent on heat shock factor in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci* 2001;**98**:7313 LP – 7318.
- Varela JA, Gethins L, Stanton C *et al.* Applications of *Kluyveromyces marxianus* in Biotechnology BT - Yeast Diversity in Human Welfare. In: Satyanarayana T, Kunze G (eds.). Singapore: Springer Singapore, 2017, 439–53.
- Verduyn C, Postma E, Scheffers WA *et al.* Effect of benzoic acid on metabolic fluxes in yeasts: A continuous-culture study on the regulation of respiration and alcoholic fermentation. *Yeast* 1992;**8**:501–17.
- Wang D, Wu D, Yang X *et al.* Transcriptomic analysis of thermotolerant yeast *Kluyveromyces marxianus* in multiple inhibitors tolerance. *RSC Adv* 2018;**8**:14177–92.
- Warner JR. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci* 1999;**24**:437–40.

## Chapter 5 – General Discussion and Future Research

The main objective of this thesis was to develop and apply ribosome profiling protocol for *K. marxianus*, a non-conventional yeast with a broad range of physiological traits that make it attractive for both food and industrial use (Fonseca *et al.* 2008; Lane and Morrissey 2010; Morrissey *et al.* 2015; Varela *et al.* 2017; Karim, Gerliani and Aider 2020). The majority of studies in *K. marxianus* have primarily focused on transcriptomic techniques (RNA-seq) to study global gene expression responses to a particular environment; such as industrial stresses (reviewed in Ha-Tran, Nguyen and Huang 2020), thus ignoring potential post-transcriptional regulation. Ribosome profiling is a technique which provides the locations of translating ribosomes *in vivo* (Ingolia *et al.* 2009), providing information on global translation which could be a useful tool to study industrial stresses or provide information on genes responsible for growth relating to particular substrates in *K. marxianus*.

In Chapter 2 of this thesis, the development of a ribosome profiling protocol is described whereby using a combination of sucrose gradient ultracentrifugation to isolate monosomes (Ingolia *et al.* 2009) and cDNA library generation techniques (McGlinchey and Ingolia 2017) previously used for ribosome profiling in *S. cerevisiae*, a protocol for *K. marxianus* was generated. It was shown that this protocol provides accurate and reliable translation information with a high triplet periodicity signal and the vast majority of RPFs are located within CDS boundaries as expected. In the future, this protocol may also be suitable for generating ribosome profiling data in other NCY such as *K. lactis* or *Y. lipolytica*.

However, as ribosome profiling is a useful technique which could be employed to study other NCYs, it has a number of caveats. Firstly, it is technically challenging and requires specialised equipment including vacuum filtration apparatus, cryo-mills and ultracentrifuges. For yeast, it is crucial to rapidly filter and flash freeze cell pellets from liquid medium to avoid any distortions that can occur with translation, in the protocol presented in Chapter 2, this process takes ~10 seconds. If rapid filtration cannot be performed, cultures can be pre-treated with cycloheximide to preserve translation but biases of pre-treatment can arise (Husmann *et al.* 2015; Duncan and Mata 2017; Santos *et al.* 2019; Sharma *et al.* 2021). Lysis is technically challenging for yeast ribosome profiling experiments, as yeast cell lysis requires cryo-milling due to the presence of a cell wall. Size selection of RPFs presents

another technical issue as size selection and gel cutting can vary between studies, and larger size selections can introduce more rRNA contaminants. rRNA contamination can vary between species and may require trial ribosome profiling experiments and carefully designed depletion oligos to reduce rRNA contaminants in the cDNA libraries. As short read lengths (~28 nt) are generated from RPFs, ambiguous mapping can be an issue as discussed in Chapter 2. The ratio of ambiguously mapped reads can vary between species, and a higher ratio of ambiguous reads can be expected in post-WGD species (due to a larger proportion of paralogous genes sharing sequences with high similarity) including *S. cerevisiae* compared to a pre-WGD species such as *K. marxianus*, giving these pre-WGD yeasts an advantage. These considerations should be taken into account when developing ribosome profiling in other yeast species.

During the analysis of the first ribosome profiling data presented in Chapter 2 and with the addition of the *K. marxianus* reference genome to GWIPS-Viz (Michel *et al.* 2014, 2018), we observed the presence of a large number RPFs in unannotated ORFs which suggested the genome annotation of the reference genome was incomplete. As the initial data was lacking in sufficient coverage for reannotation, it was decided to set up a larger experiment to test the protocol and eventually study the transcriptome of *K. marxianus*. As thermotolerance is an interesting feature in *K. marxianus*, this yeast was subjected to a rapid heat shock from 30°C to 40°C and cells were collected at 5, 15, 30 and 60 minute timepoints post heat-shock. These data were individually analysed and discussed in Chapter 4. The advantage of generating ribosome profiling data from multiple conditions allows the potential to observe and annotate genes that have no/low expression in one condition and are highly expressed in another condition, providing a more complete coverage of translation under diverse conditions. In Chapter 3, we used these new ribosome profiling data and a method heavily reliant on the genome browser GWIPS-Viz to identify and annotate genes absent from the previous genome annotation, resulting in the addition of over 170 protein-coding genes (Chapter 4). It is interesting to note that these included both novel genes (unique to *K. marxianus*) and conserved genes with homologs present in other species such as *S. cerevisiae*, suggesting current annotation techniques are not sufficient to produce a complete de novo genome annotation. This is especially important for genes unique to the *K. marxianus* genome as comparative genomic tools are more likely to fail to detect such genes. Therefore, using a combination of traditional annotation tools and ribosome profiling data is necessary to

produce a more complete and accurate genome annotation. The updated genome annotation presented in Chapter 3 may also serve as a sufficient template for annotation of other *Kluyveromyces* species; as it likely provides a more accurate model for genome annotation than using a more distant yeast model such as *S. cerevisiae*. In addition to novel gene annotation, multiple errors in the previous genome annotation such as start codon, intron and exon coordinates were corrected. In the future (as discussed in Chapter 3), additional ribosome profiling data from alternative conditions to the conditions presented in this work may allow the discovery of further additional genes, for example, a gene required for utilization of a different carbon may have been missed from our data as none of the RNA-seq or Ribo-seq samples were collected from such media. In addition, we observed the presence of many annotated ORFs in the reference genome which may not produce functional proteins due to a lack of RPFs. Therefore, an effort could be made to study these “junk” genes to uncover whether they may be non-coding RNAs, if orthologs exist in other species or as more ribosome profiling data becomes available, analyse these loci for RPFs. A more comprehensive and accurate genome annotation will allow for better genomic studies, especially in the area of novel genes, many of which can be termed as evolutionary young genes and have been shown to be differentially expressed (Doughty *et al.* 2020), one such example in *K. marxianus* has been shown to play a vital role in growth at high temperatures (Montini *et al.* 2022).

In addition to the ribosome profiling data generated, transcriptomic techniques were also employed in the “multiomic” approach discussed in Chapter 3. In the initial RNA-seq experiments that accompanied ribosome profiling, polyA selection to enrich mRNAs generated the distinct 3’ bias often seen in RNA-seq studies. In this 3’ bias, RNA-seq reads near the polyA tail had greater density than those upstream. As these reads are heavily enriched at the polyadenylation site and the polyA tail, clipping the polyA tail from these sequences revealed polyadenylation sites genome-wide, providing information on the 3’ boundaries of mRNAs. Publicly available TSS data was analysed to provide the coordinates of 5’ ends of mRNAs (Lertwattanasakul *et al.* 2015). Ribosome profiling, TSS, RNA-seq and PAS allowed a multiomic approach to study *K. marxianus* and with the use of genome browsers, it was possible to study a genomic loci in terms of transcription and translation simultaneously. An example whereby multiomic techniques were crucial in deciphering complex events included the locus of KLMX\_30357. Originally annotated as two

separate protein coding genes, both ribosome profiling and transcriptomic data would show this locus belongs to a single gene utilizing +1 frameshifting to translate the full length product. For this example, either techniques alone would have been insufficient to determine the structure of this gene.

In the future, alternative transcriptomic techniques could be employed to study other yeasts alongside ribosome profiling. An advantage of transcriptomic techniques is the wide range of techniques available to reveal 5' and/or 3' mRNA boundaries. While information on yeast TSS of multiple species is present on the yeast TSS database, species including *K. marxianus* are absent but other NCY yeasts are included such as *Y. lipolytica* and *Lachancea L. kluyveri* (McMillan *et al.* 2019). For short read sequencing, TSS-Seq (used in this study), TL-Seq, TIF-Seq, PAS-Seq and nanocage protocols are publicly available (Salimullah *et al.* 2011; Shepard *et al.* 2011; Arribere and Gilbert 2013; Pelechano *et al.* 2014). Oxford Nanopore Technologies uses protein nanopores to sequence full length mRNAs (via direct RNA-sequencing) or cDNAs, allowing single reads to reveal the full length of an mRNA (Garalde *et al.* 2018). Direct RNA-seq is advantageous as it negates potential biases of cDNA library generation and allows detection of RNA modifications (m6A) which may have regulatory functions (Liu *et al.* 2019).

N-terminal proteoforms were also discovered and using multiomic data it was possible to determine how these proteoforms are produced, either from leaky scanning of the first start codon (AUG or non-AUG) or transcriptionally derived via alternative TSS. Many of these NTEs were shown to harbour MTS sequences. One such example identified from these analyses was *BAT1*, which was later explored in an independent follow-up study (Coral-Medina *et al.* 2022). In *S. cerevisiae*, the genome encodes two paralogs of *BAT1* due to a whole genome duplication or hybridization event (Wolfe and Shields 1997; Marcet-Houben and Gabaldón 2015). *BAT1* encodes the mitochondrial isoform (via mitochondrial targeting signal at the N-terminus) while *BAT2* encodes the cytosolic isoforms. In Coral-Medina *et al.*, 2022, it was shown that *K. marxianus*, which has only one *BAT1* gene, uses two mRNA isoforms with the shorter mRNA encoding cytosolic Bat1 and the longer mRNA encoding mitochondrial Bat1. In the future, for both *K. marxianus* and other species, there are likely more cases of such gene-level regulation which could be discovered using ribosome profiling

and/or transcriptomic techniques. These data will allow researchers to go beyond comparative genomics and study how genes have evolved different or similar mechanisms of regulation. A surprising result in Chapter 3 was the discovery of antisense mRNAs that are translated. Such cases were first reported in the fission yeast *S. pombe* using ribosome profiling (Duncan and Mata 2014). These antisense mRNAs will require future work to determine if these encode functional proteins, work as regulators of sense mRNAs or other mRNAs. Efforts could be made to study these antisense mRNAs in other yeasts to find similarities as well as carefully designed functional studies with mutants. Bioinformatic studies could be used to identify such antisense mRNAs in *S. cerevisiae* or more closely related species such as *K. lactis*, these could potentially include antisense mRNAs on the orthologous genes which may suggest conservation or other non-related genes. For functional studies, the promoter region of the antisense mRNA could be mutated to prevent transcription or expression of the antisense mRNA. Alternatively, an expression vector to overexpress the antisense mRNA could provide phenotypic results. Ribosome profiling and RNA-seq could also be used to study how these antisense mRNAs hybridize to the sense mRNA and affect mRNA translation or mRNA stability.

In Chapter 4, we used ribosome profiling and RNA-seq to study how *K. marxianus* adapts to a rapid increase in temperature (30°C to 40°C). The most surprising result is that at 5 minutes post heat-shock, *K. marxianus* upregulates aerobic respiration along with expected biological processes such as the unfolded protein and heat shock response. Moreover, we show disparity in how nuclear and mitochondrially encoded genes whose encoded proteins belong to the same complex (such as the ATPase synthase complex) are upregulated differently. Deciphering the network of this signalling between mitochondria and nuclear regulation will require further work to identify the signal transduction pathways involved. While differential gene expression analysis strongly suggests an increase in cellular respiration, a more sensitive and rapid method will need to be employed to study mitochondrial ATP production (mitochondrial membrane potential) and/or cellular oxygen uptake assays.

In conclusion, this work demonstrates the usefulness of ribosome profiling and transcriptomic techniques in deciphering both the transcriptional and translation landscape of *K. marxianus*, both expanding the known protein-coding repertoire, mechanisms of mRNA translation and

potential regulation of genes. These methods will be a valuable tool for future studies in both *K. marxianus* (as presented in this work) and other NYCs.

## References

- Arribere JA, Gilbert W V. Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res* 2013;**23**:977–87.
- Coral-Medina A, Fenton DA, Varela J *et al.* Regulated use of alternative Transcription Start Sites controls the production of cytosolic or mitochondrial forms of branched-chain aminotransferase in *Kluyveromyces marxianus*; *bioRxiv* 2022:2022.04.27.489738.
- Doughty TW, Domenzain I, Millan-Oropeza A *et al.* Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts. *Nat Commun* 2020;**11**:2144.
- Duncan CDS, Mata J. The translational landscape of fission-yeast meiosis and sporulation. *Nat Struct Mol Biol* 2014;**21**:641–7.
- Duncan CDS, Mata J. Effects of cycloheximide on the interpretation of ribosome profiling experiments in *Schizosaccharomyces pombe*. *Sci Rep* 2017;**7**:10331.
- Fonseca GG, Heinzle E, Wittmann C *et al.* The yeast *Kluyveromyces marxianus* and its biotechnological potential. *Appl Microbiol Biotechnol* 2008;**79**:339–54.
- Garalde DR, Snell EA, Jachimowicz D *et al.* Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* 2018;**15**:201–6.
- Ha-Tran DM, Nguyen TT, Huang C-C. *Kluyveromyces marxianus*: Current State of Omics Studies, Strain Improvement Strategy and Potential Industrial Implementation. *Ferment* 2020;**6**, DOI: 10.3390/fermentation6040124.
- Hussmann JA, Patchett S, Johnson A *et al.* Understanding Biases in Ribosome Profiling Experiments Reveals Signatures of Translation Dynamics in Yeast. *PLoS Genet* 2015;**11**:e1005732.
- Ingolia NT, Ghaemmighami S, Newman JRS *et al.* Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 2009;**324**:218–23.
- Karim A, Gerliani N, Aider M. *Kluyveromyces marxianus*: An emerging yeast cell factory for applications in food and biotechnology. *Int J Food Microbiol* 2020;**333**:108818.
- Lane MM, Morrissey JP. *Kluyveromyces marxianus*: A yeast emerging from its sister's shadow. *Fungal Biol Rev* 2010;**24**:17–26.

- Lertwattanasakul N, Kosaka T, Hosoyama A *et al*. Genetic basis of the highly efficient yeast *Kluyveromyces marxianus*: complete genome sequence and transcriptome analyses. *Biotechnol Biofuels* 2015;**8**:47.
- Liu H, Begik O, Lucas MC *et al*. Accurate detection of m6A RNA modifications in native RNA sequences. *Nat Commun* 2019;**10**:4079.
- Marcet-Houben M, Gabaldón T. Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLOS Biol* 2015;**13**:e1002220.
- McGlinchy NJ, Ingolia NT. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* 2017;**126**:112–29.
- McMillan J, Lu Z, Rodriguez JS *et al*. YeasTSS: an integrative web database of yeast transcription start sites. *Database (Oxford)* 2019;**2019**:baz048.
- Michel AM, Fox G, M Kiran A *et al*. GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Res* 2014;**42**:D859–64.
- Michel AM, Kiniry SJ, O'Connor PBF *et al*. GWIPS-viz: 2018 update. *Nucleic Acids Res* 2018;**46**:D823–30.
- Montini N, Doughty TW, Domenzain I *et al*. Identification of a novel gene required for competitive growth at high temperature in the thermotolerant yeast *Kluyveromyces marxianus*. *Microbiology* 2022;**168**, DOI: 10.1099/mic.0.001148.
- Morrissey JP, Etschmann MMW, Schrader J *et al*. Cell factory applications of the yeast *Kluyveromyces marxianus* for the biotechnological production of natural flavour and fragrance molecules. *Yeast* 2015;**32**:3–16.
- Pelechano V, Wei W, Jakob P *et al*. Genome-wide identification of transcript start and end sites by transcript isoform sequencing. *Nat Protoc* 2014;**9**:1740–59.
- Salimullah M, Sakai M, Plessy C *et al*. NanoCAGE: a high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harb Protoc* 2011;**2011**:pdb.prot5559-pdb.prot5559.
- Santos DA, Shi L, Tu BP *et al*. Cycloheximide can distort measurements of mRNA levels and translation efficiency. *Nucleic Acids Res* 2019;**47**:4974–85.
- Sharma P, Wu J, Nilges BS *et al*. Humans and other commonly used model organisms are resistant to cycloheximide-mediated biases in ribosome profiling experiments. *Nat Commun* 2021;**12**:5094.
- Shepard PJ, Choi E-A, Lu J *et al*. Complex and dynamic landscape of RNA polyadenylation



revealed by PAS-Seq. *RNA* 2011;**17**:761–72.

Varela JA, Gethins L, Stanton C *et al.* Applications of *Kluyveromyces marxianus* in Biotechnology BT - Yeast Diversity in Human Welfare. In: Satyanarayana T, Kunze G (eds.). Singapore: Springer Singapore, 2017, 439–53.

Wolfe KH, Shields DC. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 1997;**387**:708–13.