Gene expression differences and functional diversification of glycolytic gene paralogs in *Saccharomyces cerevisiae*

Nicholas Brown

Wittkopp Lab – University of Michigan

1 April 2023

A thesis submitted in partial fulfillment of the Degree of Bachelor of Science in Microbiology with Honors

Advisor: Dr. Patricia Wittkopp, Ph.D.
Reader #1: Dr. Timothy James, Ph.D.
Reader #2: Dr. Jianzhi Zhang, Ph.D.

Abstract

Genomes of all organisms are full of genes that duplicated and then subsequently diversified in function during evolution. A primary goal of evolutionary genetics is to mechanistically understand how genes evolve following duplication such that both copies become essential and retained by evolution across millions of years of evolutionary time. In my research, I used the evolution and diversification of the three *TDH* genes within *Saccharomyces cerevisiae*, to investigate this question. The *TDH* gene duplicates (*TDH1*, *TDH2*, *TDH3*) are involved in the fundamental process of glycolysis, through which sugars are converted to energy, and specifically catalyze the conversion of glyceraldehyde-3-phosphate in this pathway. My research asks explicitly if the three *TDH* genes have diverged with respect to their effects on growth rate and whether this divergence happened through changes in the gene promoter or the protein-coding region. I first measured the effect of each *TDH* gene on growth rate by using CRISPR-Cas9 to delete them individually and in combination. I then determined whether the divergence between genes was because of changes in the gene promoters or the protein-coding sequences by swapping the promoters and protein-coding sequences between the *TDH* genes and measuring growth rate. My data shows that the *TDH* duplicates have diverged in some functions but retained other conserved functions. Further, my data shows that most functional differentiation between the *TDH* genes, with respect to effects on growth rate, occurred through changes in the promoter that alters gene expression. However, I also discovered that the TDH paralogs may have roles beyond glycolysis and fermentative growth. To investigate this, fluorescent microscopy was used to determine if the subfunctionalization of *TDH* regulatory sequences over evolutionary time was followed by the neofunctionalization of protein-coding sequences.

Acknowledgments

This thesis would not have been possible without the support of many others. First, I would like to thank my advisor, Dr. Patricia Wittkopp, for her unparalleled mentorship and guidance throughout my time in the lab. I am beyond grateful for the grace and flexibility she provided me in pursuing my research. I have been a part of the Wittkopp Lab for over three years at this point, and in that time I have experienced a global pandemic, unusual academic settings, and a multitude of setbacks from failed experiments (Science, right?). Throughout it all, Trisha has been there to help me and keep me excited about the science that I am passionate about.

I also owe many thanks to Dr. Mo Siddiq for teaching me both the wet and dry lab techniques that made my investigation possible; without his attentiveness and patience, I would have lost my grip on the science quickly. Whether we were in the lab talking about science or the current state of the world, I knew that I could come to him to discuss whatever was on my mind.

Much of this thesis would also not have been possible without the help and friendship of Hannah Kania, who constantly reminded me to never sell myself short of my capabilities and to always think outside of the box. I would also like to thank past and present Wittkopp Lab members for their advice and support throughout my time in the lab: Lisa Kim, Alicia Wang, Henry Ertl, Dr. Abby Lamb, Dr. Eden McQueen, Anna Redgrave, Dr. Pétra Vande Zande, Kevin Huang, Dr. Ayse Tenger-Trolander, Holly Scheer, Dr. Homa Papoli Yazdi, Ayushi Goel, Dr. Nastassia Vlasava, and Dr. Erick Bayala Rodriguez. I truly can not imagine a more supportive space where I could have learned science.

Finally, I would like to thank Kimberly Haveman-Kudlack for instilling in me a love for science and research when I was in high school and my parents for supporting and uplifting me in my education.

Table of Contents

Introduction

Gene duplications have been essential in the evolution of most species' genomes. Gene duplication events are evolutionary phenomena by which a gene that existed in one copy is duplicated into more copies. The resulting copies are called paralogs. Initially, they cannot be differentiated from each other, but, over evolutionary time, they diverge. These duplication events then provide genetic information and material for selection, mutations, and drift to act upon (Crow 2006). Gene duplication events are highly abundant, and every bacterial, eukaryotic, and archaeal genome sequenced have paralogous genes. However, the exact mechanisms through which duplicated genes are retained over millions of years of evolutionary time are unclear (Zhang 2003). The effect of a duplication event is often that one or more of the duplicated gene pairs is silenced or lost because of the accumulation of mutations that are negatively selected for; because one copy often becomes nonfunctional, most duplication events typically do not affect the organism's evolution (Lynch 2002). However, this is not always the case, and the genomes of modern organisms provide evidence that duplicate copies are retained frequently enough to affect evolution. Some models have been constructed to explain why gene duplicates may be kept.

While gene duplication is a complicated process to study, evolutionary biologists have categorized two significant ways through which paralogs are kept following a duplication: regulatory sequence changes (Figure I.1) or protein-coding sequence changes (Figure I.2) (Hahn 2009). Within these changes, one must also look into the models through which these changes function. These models are neofunctionalization, subfunctionalization, and gene conservation. Each of these explanations for kept paralogs have different, empirically tractable predictions. By investigating each of them, we can better understand why gene duplicates are so abundant across the whole of life.

One outcome following gene duplication is neofunctionalization. Neofunctionalization is the process by which after a gene is duplicated, one of the daughter genes gains a new function either in its regulatory region or in its protein-coding region. The role of neofunctionalization in gene duplication is undoubtedly important, but because it is hard to test for the evolution of new functions, the phenomenon is not as widely understood (Hahn 2009, Siddiq et al. 2017). When thinking about how regulatory sequence changes might affect how gene duplicates are kept, we must think of changes to a paralog's promoter following a duplication event that gives rise to a new function for a paralog by changing when and how much gene product is made. In the context of protein-coding sequence changes and how that may contribute to a kept paralog, we must think about how a gene duplication event caused changes to the coding sequence, therefore giving rise to a new protein function.

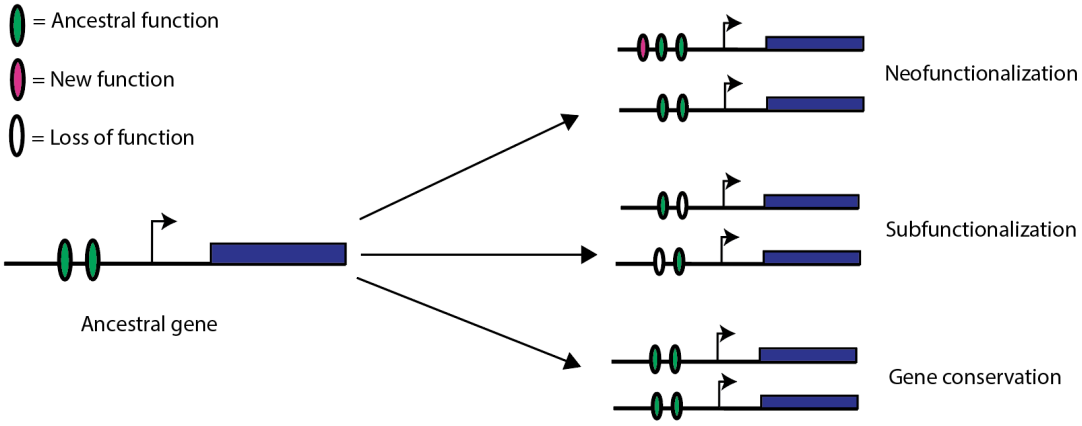

Figure I.1: Adapted from Hahn 2009. Models of how gene duplicates are kept due to changes in regulatory sequences.

Another reason that gene duplicates are kept is called subfunctionalization. Subfunctionalization is the phenomenon through which each paralog to an ancestral gene keeps some ancestral function, but reciprocally loses other functions (Lynch & Force 2000, Hahn 2009,

Voordeckers et al. 2015). Subfunctionalization can happen without any adaptive benefit, because the same functions previously specified by one gene are now specified by two genes (van Hoof 2005), or it can facilitate adaptation by allowing different functions to be optimized in different paralogs (Barkman & Zhang 2009). When thinking about how subfunctionalization might lead to regulatory changes in homologous paralogs that allow for the paralogs to be kept, we must keep in mind that *part* of the regulatory sequence is keeping ancestral function, but other redundant function in the paralog(s) may be lost. When thinking about protein-coding sequence changes that give rise to paralogous genes in the context of subfunctionalization, we must remember that *part* of the protein-coding sequence is retaining ancestral function, but other redundant functions may be lost.

A final model that helps to explain why gene duplicates are kept is that of gene conservation. This model just states that a gene duplication event happened because an organism needed to evolve more of the same gene to survive selection, potentially because of cellular demands (Hahn 2009). There are two reasons why this model might be favored: redundancy and dosage. If an ancestral gene lost some function due to a mutation, gene conservation might retain redundant function that is needed. In the case of dosage, it may be purely advantageous to have more of the same gene for an organism's survival. Evolutionary biologist Susumu Ohno made the convincing claim that if a certain necessary gene product is needed for an organism's survival, duplication of the genes that make the product carries evolutionary sense in the context of selection (1970). Therefore, by conserving the function of necessary genes in either their regulatory or protein-coding regions, an organism might have a better chance of survival.

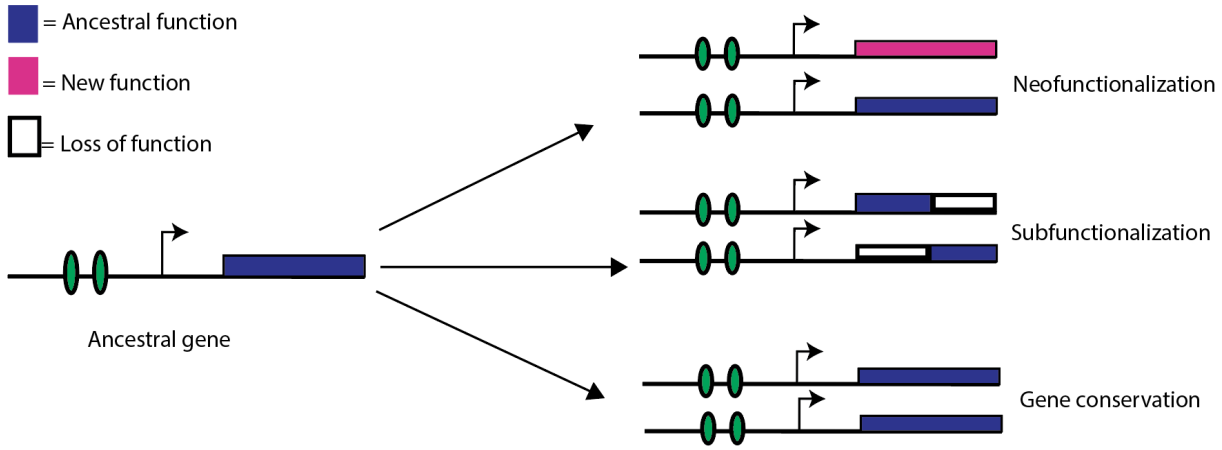

Figure I.2: Adapted from Hahn 2009. Models of how gene duplicates are kept due to changes in the protein-coding sequences.

Evolutionary biologists have long studied gene duplications and have developed theoretical rationales for how they contribute to evolution. However, there are still gaps of knowledge in the field that make it difficult to discern why duplicate genes are kept because the empirical data necessary to discern between different scenarios following gene duplication is not often available. An immensely valuable model organism that is used to help fill these gaps in knowledge is the baker's yeast, *Saccharomyces cerevisiae*. This yeast is a single-celled eukaryote with rapid generation times, and there is a large availability of molecular toolkits that allow for experimentation on the organism at the molecular and genomic levels. Along with this, some data supports that the yeast speciated following a whole-genome duplication (Boonekamp et al. 2018), making it ideal for studying duplication events because the yeast have a multitude of gene duplicates in their genome and the gene duplications have been a fundamental part of their evolutionary history. Interestingly, many housekeeping genes that are critical for yeast cell metabolism arose from gene duplication events (Boonekamp 2018), and the understanding as to why so many of the gene copies were kept is minimal. All of these things combined make *S.*

cerevisiae a great model organism for studying gene duplication and why paralogous genes that remain in their genome are evolutionarily conserved.

This thesis attempts to provide an understanding as to why metabolic gene paralogs in *S. cerevisiae* were kept following a gene duplication event, and whether the changes that allowed for the conservation of the paralogs were due to changes in the regulatory or protein-coding sequences of the genes. Investigation into this was done using CRISPR-Cas9 genome editing technology and lab-generated strains of *S. cerevisiae*. The first part of this thesis explores the changes to the individual *TDH* paralogs to explain why they are conserved in the yeast genome. The second part of this thesis examines novel functions that may have arisen in yeast due to such changes.

Evolution of *TDH* expression differences in *S. cerevisiae*

Introduction

The ability to grow and metabolize nutrients is arguably the most important function an organism has in ensuring its survival. For organisms to grow, they must perform glycolysis–a metabolic reaction found across all living organisms. The process of glycolysis is fundamental and simple: available sugars in the form of glucose are broken down to obtain ATP for energy. Glycolysis is nearly the same across each domain of life (Archaea, Bacteria, and Eukarya), and yet different organisms often have slightly different sets of highly conserved metabolic genes that keep the process running smoothly. One organism that serves as a model for how glycolysis works is the baker's yeast, *Saccharomyces cerevisiae*. The yeast is incredibly efficient in its glycolytic ability; however, it is strange in the fact that even in the presence of oxygen, it chooses to take the fermentation route to produce energy (Gonçalves & Planta 1998), even though aerobic respiration allows more energy to be produced per unit of glucose. One potential reason is because glycolysis is faster and allows more rapid production of energy from sugar. The genes that encode for steps in the glycolytic and

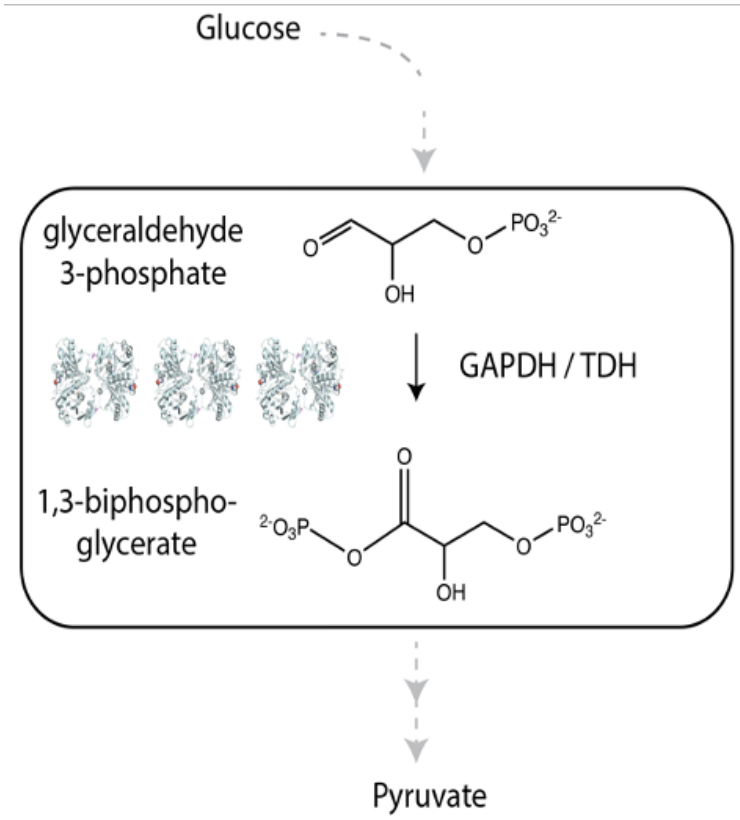

Figure 1.1: TDH converts glyceraldehyde 3-phosphate to 1,3-bisphosphoglycerate during the sixth step of glycolysis.

fermentative processes in *S. cerevisiae* are highly conserved (Bonnekamp et al. 2018), important for the organism's natural history, and serve as invaluable tools to study how and why the yeast functions in the way that it does.

S. cerevisiae belongs to a group of species that evolved following a whole-genome duplication, and with that, their genome has an abundance of paralogous genes with potentially redundant functions. Metabolic genes involved in glycolysis are some of the many that duplicated (Boonekamp et al. 2018). One example is the *TDH* genes. *TDH*s are homologous to the human *GAPDH*, which encodes the enzyme glyceraldehyde-3-phosphate dehydrogenase. *S. cerevisiae* has three *TDH* paralogs called *TDH1, TDH2,* and *TDH3*. TDH proteins catalyze the conversion of glyceraldehyde 3-phosphate to 1,3-bisphosphoglycerate in the sixth step of glycolysis. The contribution of each TDH gene to glycolytic activity varies, and the first studies of the genes found that *TDH1* contributes 10 to 15% activity, *TDH2* contributes 25 to 30% activity, and *TDH3* contributes 50 to 60% activity (McAlister & Holland 1985). These data show that the *TDH* genes contribute differently, but why they do so is unknown. For example, the *TDH* genes may have diverged in their protein-coding sequence, in their expression levels, or both. When scientists were making these discoveries, it was hard to use molecular genetics to test the different models for how the different *TDH* genes diverged during evolution because the resources were not available. These knowledge gaps, however, can be overcome because of the emergence of next-generation techniques like CRISPR-Cas9 genome editing (Doudna & Charpentier 2014). I took advantage of this technology to test my research questions in *S. cerevisiae*.

The primary goal of this chapter is to understand *why* the *TDH* paralogs in *S. cerevisiae* contribute differently to glycolysis and growth. It is known that an ancestral *TDH* gene underwent a gene duplication event that gave rise to the three paralogs; my research question asks whether

the paralogs functionally diverged because of changes in their regulatory sequences or changes in their protein-coding sequences. To test this, I used CRISPR-Cas9 to delete *TDH* protein-coding sequences individually and in parallel to measure each gene's contribution to growth. Next, I tested whether the difference in each gene's contribution was caused by changes in the promoter that controls regulation of each *TDH* gene or in the protein-coding sequence by mixing and matching promoter-protein regions. Following this, I conducted a growth rate analysis with a 96-well microplate reader to test the consequences of these swaps and deletions.

Materials and Methods

Guide RNA (gRNA) design and cloning. I designed CRISPR-Cas9 plasmids to create targeted double-strand breaks in different *TDH* genes using the Laughery et al. 2015 strategy. gRNAs were cloned into pML-104 plasmids via Gibson assembly. To find suitable gRNA target sequences in the *S. cerevisiae* genome, Benchling's gRNA design tool was used with specificity for the CRISPR-Cas9 system. The sequences of primers that were used to create gRNAs for cloning into pML-104 are listed in Table 2 (Appendix). Once primers were obtained to hybridize gRNAs, a PCR reaction was conducted so that the gRNA primers could hybridize. To clone gRNAs into pML104, a bacterial glycerol stock with the pML104 plasmid, was first grown in 5 mL LB broth at 37°C overnight. Following overnight growth of the plasmid, the plasmid was isolated by conducting a Quigen Mini-Prep using the instructions provided in the kit. After the plasmid was isolated and the concentration checked using a Thermo Fisher Scientific NanoDrop spectrophotometer, the plasmid was digested with BclI and SwaI restriction enzymes so that the plasmid was open for cloning. The digested plasmid was cleaned and concentrated using NEB's DNA Clean and Concentrate kit following the instructions provided in the kit. Digestion was confirmed using gel electrophoresis. Following this, hybridized gRNAs were ligated into digested

pML104. Approximately 100 ng of digested plasmid was used for the ligation, along with 1/20 diluted hybridized gRNA. Following ligation of the plasmid, the plasmid was transformed into XL-10 Gold *E. coli* competent cells; cells with the transformed plasmid were then glycerol stocked in a 20% glycerol solution and stored in a -80℃ freezer. The full procedure is linked in the Appendix.

CRISPR-Cas9 repair fragment design and cloning. Repair fragments for each CRISPR-Cas9 experiment were designed concerning the region in the yeast genome that they would be integrated into. There were two kinds of repair fragments that were needed: repair fragments for deletion strains and repair fragments for chimeric strains. Each fragment had both 5' and 3' homology arms to ensure complementarity when integrated into the target areas; without complementarity, there is no guarantee that the fragments will correctly integrate into the yeast genome. Repair fragments were Gibson assembled into a plasmid with the appropriate homology arms for the locus that was being targeted. The sequences of the primers that were used to create repair fragments are listed in Table 3 (Appendix).

Plasmid prepping for obtaining gRNAs and CRISPR-Cas9 repair fragments. Plasmids that contained CRISPR-Cas9 gRNAs or CRISPR-Cas9 repair fragments were isolated from XL-10 *E. coli* cells using a Qiagen Plasmid Mini-Prep Kit. To isolate the plasmids, instructions listed within the mini-prep kit were followed. Isolated plasmid concentrations were obtained using a Thermo Fisher Scientific NanoDrop Spectrophotometer. Repair fragments were amplified from plasmids using Thermo Fisher DreamTaq polymerase and primers per manufacturer instructions.

***CRISPR-Cas9 genome editing of* S. cerevisiae.** To create the knockout strains and the chimeric promoter/protein-coding sequence strains of *S. cerevisiae*, a CRISPR-Cas9 genome editing protocol adapted from Laughery et al. 2015 was used. Wild-type *S. cerevisiae* + Δ*URA3* was grown up in 5 mL of YPD overnight at 30℃, and diluted to OD of 0.75[1] before the experiments were conducted. CRISPR-Cas9 yeast transfections were done in which gRNAs targeted to the C-terminus of *TDH* protein-coding sequences were used with Cas9 to make a double-strand break at the C-termini to either delete the *TDH* gene (Figure 1.2) or recombine in a non-native *TDH* gene to create a *TDH* promoter/protein-coding sequence chimera (Figure 1.3). gRNAs were stored in pML104 plasmids that also included *URA3* so as to allow transfected colony growth on selective media. Following CRISPR-Cas9 transfections, transfected yeast were plated onto SC-URA agar plates to select for successful transfectants. Plates were incubated for two days. Following this period, the plates were observed for successful transfectants. Observed colonies were moved to 5'-FOA plates to remove residual genome-editing plasmids from the colonies and select for the absence of *URA3*. These plates were incubated for 2 days at 30℃. Following the incubation period, colonies were checked for the presence of the correct genotype by PCR amplification and gel electrophoresis. Successful transfectants were then sent to Eurofins Genomics for sequencing. With the confirmation of successful transfectants, new yeast strains were glycerol stocked in 20% glycerol and placed into a -80℃ freezer for storage. Strains are represented in Table 1 (Appendix). The full procedure is linked in the Appendix.

[1] An OD measure of 0.75 indicated the stage of yeast growth called the log phase. This phase of growth is the most robust, and it is the stage of growth most efficient for conducting transfections.

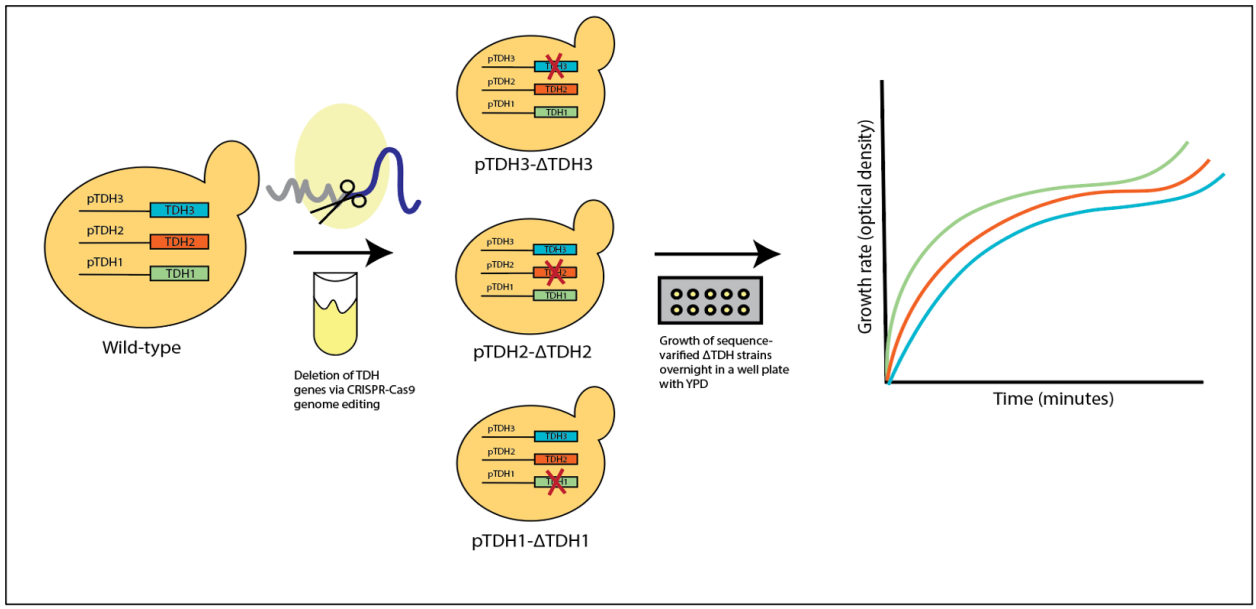

Figure 1.2: Deletion of *TDH* genes individually and combinatorially. *TDH* genes were deleted individually and in combination using CRISPR-Cas9 genome editing. Δ*TDH* gene repair fragments were designed and assembled into a plasmid with homology arms directed to the different *TDH* loci via Gibson assembly. The deletion constructs were then used in the CRISPR-Cas9 experiment to integrate them into the native loci. Δ*TDH* strains were sequence verified. The sequence verified strains were then grown up in YPD growth media overnight in a 96-well plate. Growth rate of each strain was estimated by monitoring optical density with a plate reader at 20-minute intervals.

Growth analysis of deletion and chimeric* TDH *gene strains. The constructed deletion and recombinant strains were assayed for growth. Fitness was measured as a function of growth rate. To quantify growth, yeast strains were inoculated in 5 mL of YPD and incubated on a culture tube shaker at 30°C overnight. Following the incubation, 1uL of the deletion yeast strains (Figure 1.2) and chimeric yeast strains (Figure 1.3) were individually inoculated in 5-6 wells of 100 uL of YPD

in a 96-well microplate. Growth in each well was measured in a plate reader, which took OD readings every 20 minutes for forty-eight hours to estimate strain fitness as a measure of growth.

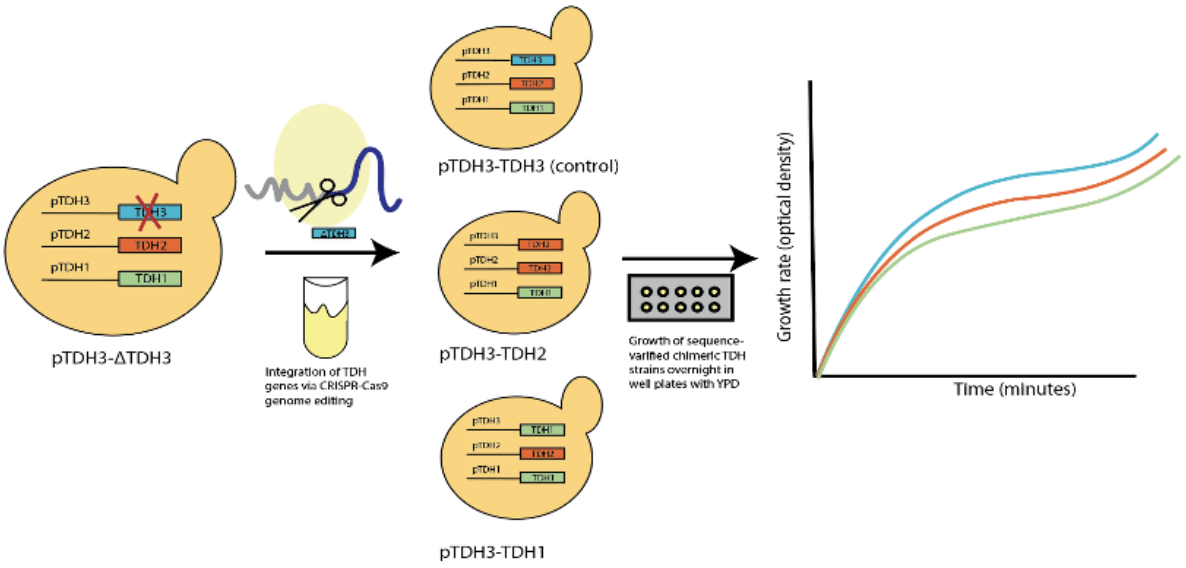

Figure 1.3: Creation of chimeric *TDH* promoter/protein-coding sequence strains of *S. cerevisiae*. *TDH* genes were swapped into the native *TDH3* locus to test promoter and protein coding regions' functions in fitness using CRISPR-Cas9 genome editing. *TDH* gene repair fragments were cloned from a wild-type strain and amplified using PCR methods. Different promoter-protein combinations were assembled into a plasmid with homology arms directed to the *TDH3* locus via Gibson assembly. These chimeric repair fragments were then utilized in the CRISPR-Cas9 experiment to integrate them into the native *TDH3* locus. *TDH* chimeric strains were sequence verified. The sequence verified strains were then grown up in YPD growth media overnight in a 96-well plate. Growth rate of each strain was estimated by monitoring optical density with a plate reader at 20-minute intervals.

Statistical analysis. All statistical analyses were performed in RStudio. For each replicate of each strain, the OD600 measurements were first normalized. To estimate growth parameters, the OD600 data were fitted with the Gompertz growth function using the nonlinear least squares method, as implemented in R (stats package). The maximum rate of increase was estimated as a part of the

fitted-model. The mean and confidence intervals for each strain was estimated from the fitted values for the replicates measured on that day.

Results

TDH3 *is required for wild-type growth of* S. cerevisiae

To determine the contribution of each *TDH* paralog to yeast fitness, gene knockouts were conducted individually and in parallel with CRISPR-Cas9 genome editing technology (adapted from Laughery et al. 2015). The fitness effect of gene knockouts was tested using a 96-well microplate reader. Deletion yeast strains (Table 1, Appendix) were grown in 5 mL of YPD overnight at 30°C and then inoculated into 100 uL of YPD in 5-6 wells of a 96-well microplate. Fitness as an effect of growth was measured as an optical density (OD) reading every twenty minutes for forty-eight hours. Normalized growth curves of deletions strains are presented in Figure 1.4 and statistical analysis of growth curves is illustrated in Figure 1.5. The assay revealed that when *TDH3* is deleted, there is a significant drop in fitness. When *TDH2* is deleted, there is no significant decrease in growth rate. Similarly, when *TDH1* is deleted, there is also no effect on organismal fitness. When *TDH1* and *TDH2* are deleted in combination, there is not a significant effect on growth. In contrast, the Δ*TDH1* + Δ*TDH3* displayed decreased fitness relative to

wildtype, and the Δ*TDH2* + Δ*TDH3* strain saw an even greater drop in fitness than the Δ*TDH3* genotype alone.

P_{TDH3}*'s disproportionate contribution of protein drives fitness in* S. cerevisiae

Chimeric promoter/protein-coding sequence strains of *S. cerevisiae* were constructed in a Δ*TDH3* background (Table 1, Appendix). I kept the *TDH3* promoter the same and changed the protein-coding region to test whether the protein-coding regions were interchangeable with respect to their effect on growth. These strains were grown in 5 mL of YPD overnight at 30℃ and then inoculated in YPD in 5-6 wells of a 96-well microplate. Fitness as an effect of growth was measured as an optical density (OD) reading every twenty minutes for forty-eight hours.

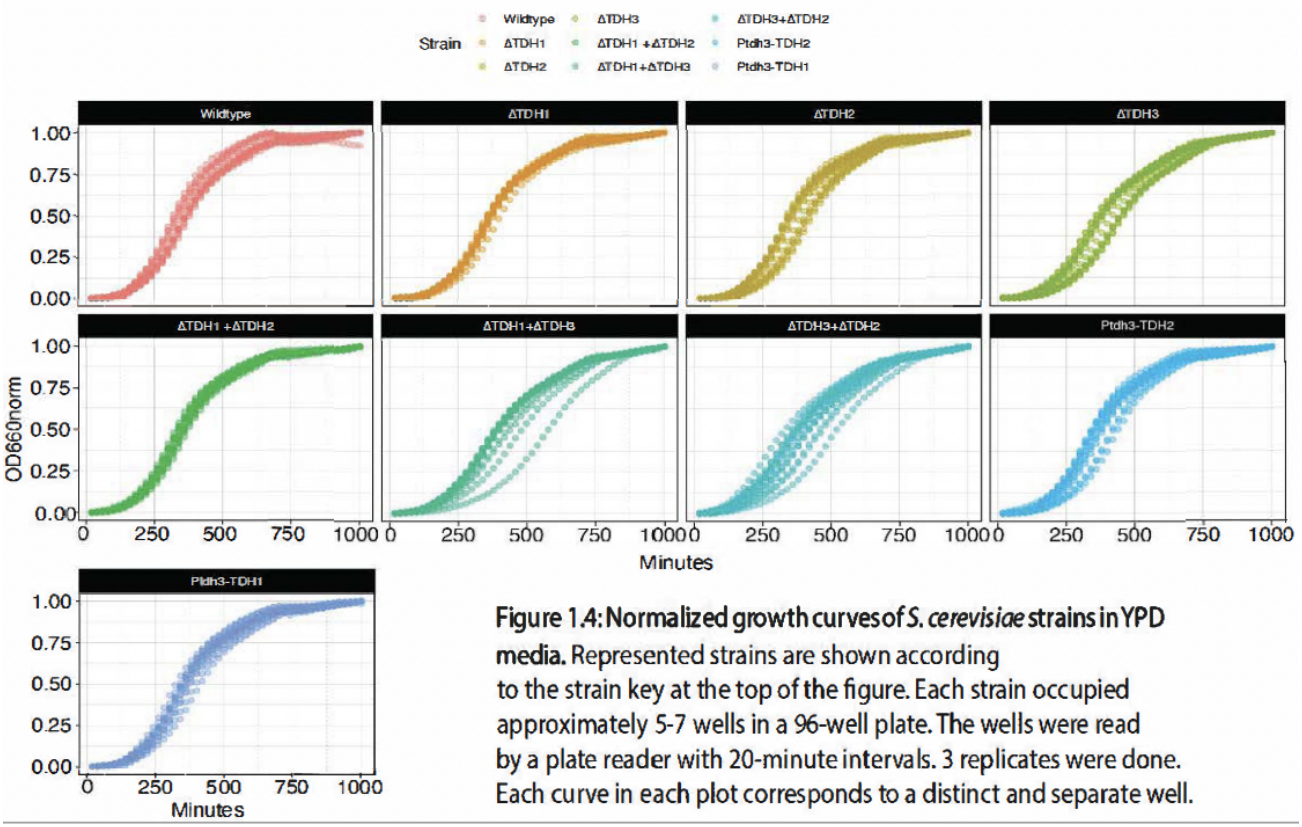

Figure 1.4: Normalized growth curves of *S. cerevisiae* strains in YPD media. Represented strains are shown according to the strain key at the top of the figure. Each strain occupied approximately 5-7 wells in a 96-well plate. The wells were read by a plate reader with 20-minute intervals. 3 replicates were done. Each curve in each plot corresponds to a distinct and separate well.

Normalized growth curves of chimeric strains are presented in Figure 1.4 and statistical analysis of the growth data is shown in Figure 1.5. The assay revealed that when the *TDH3* promoter

(P_{TDH3}) is driving the expression of *TDH1* and *TDH2* at the native *TDH3* locus, a wild-type growth phenotype is observed.

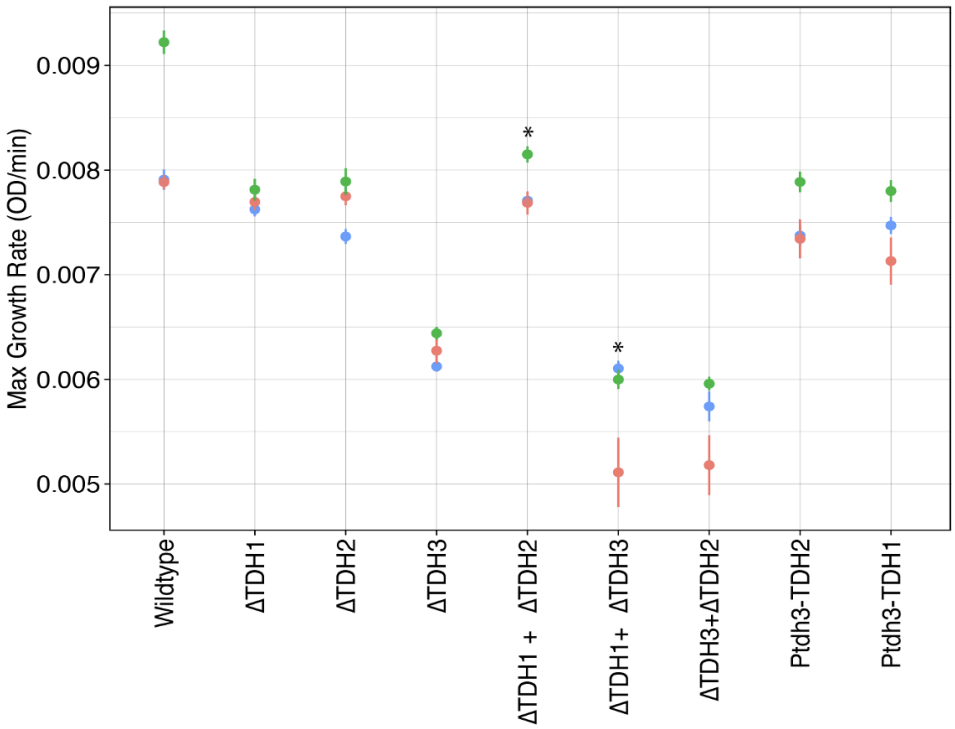

Figure 1.5: Statistical analysis of normalized growth curves. The data shown represents growth assays from 3 different days as 3 different colors. Each dot represents the mean growth rate for each day. Error bars represent 95% confidence intervals from each day. Data analysis was done in R.

Discussion

The field of evolutionary genetics has long been asking the question of how the evolution of genome structure allows for the speciation and adaptation of life. The phenomenon of gene duplication is especially interesting to evolutionary geneticists because it happens in every kingdom of life and it explains how genes or whole genomes have evolved to survive selection, therefore contributing to the evolution of organisms. Using paralogous genes to study how and

why gene duplicates are kept after a duplication event is particularly useful when the paralogs are accessible for experimentation. Eukaryotic microbes offer an interesting model system to study gene duplication because they are empirically tractable. Here, I sought to elucidate how highly conserved but differentially expressed metabolic gene paralogs *TDH1, TDH2,* and *TDH3* in *S. cerevisiae* have been kept following millions of years of evolution. In particular, I investigated whether they evolutionarily diverged because the encoded proteins are differentially expressed, because the proteins have different functions

The results of my investigation into the conservation of *TDH* paralogs led to the conclusion that *TDH3* is the gene most necessary for the fitness of *S. cerevisiae* and wild-type growth. When it is deleted, whether individually or in combination with other *TDH* paralogs, the fitness of the yeast decreases under standard growth conditions. Furthermore, it is sufficient for growth when it is the only functioning paralog; when *TDH1* and *TDH2* are both deleted, the organism's growth is not significantly affected. To take this further, when *TDH1* and *TDH2* are driven by the *TDH3* promoter,, in a *TDH3*-null background, the promoter highly expresses the other paralogs. The other paralogs, when expressed under the control of P_{TDH3}, rescue the growth defects of the *TDH3* deletion. Therefore, it can be inferred that there is differential expression of the *TDH* paralogs because changes in the regulatory sequences of the paralogs function to maintain them in the *S. cerevisiae* genome.

There are limitations to the work that was done for this chapter of this thesis. One large consideration is that phenotyping of many yeast strains was not able to be conducted because they were not engineered at the time that growth assays were performed. Only chimeric strains with P_{TDH3} driving expression at the native *TDH3* locus were phenotype, which in turn limits the conclusions that can be made about whether or not the *TDH3* regulatory sequence fully contributes

to fitness. Since phenotyping took place, chimeric strains in which P_{TDH2} and P_{TDH1} are driving the expression of each *TDH* protein-coding sequence at their respective loci have been engineered, but they have not been phenotyped. Conducting these growth assays now would give a better understanding of how the regulatory sequences of the other *TDH* genes may be contributing to the proposed conservation and divergence in the function of these paralogs. This being said, one can still use models of how gene duplicates are kept to make sense of the growth data that was collected for this thesis.

An explanation for the observed differences in how the *TDH* genes are regulated and kept could be explained by the phenomenon of subfunctionalization. Lan and Pritchard (2016) wrote that the down-regulation of certain gene duplicates, which is observed in regards to *TDH1* and *TDH2*, functions to maintain duplicate genes as it allows for the slow subfunctionalization of said genes. All three of the genes' protein products seem to have retained some ancestral glycolytic function based on the results presented in this thesis, but it is also apparent that each *TDH* protein product does not contribute equally to the phenotype of cell growth. If *TDH1* and *TDH2* were down-regulated following the duplication event that gave rise to them, they would not have been silenced but only turned on when only *TDH3* was not sufficient for growth. Also, subfunctionalization could have happened with respect to *when* the three genes are expressed. For example, I only looked at the exponential growth phase in yeast and I saw that *TDH3* is most needed for growth during this phase; the other *TDH* genes may be expressed more highly during respiratory growth or gluconeogenesis, when the yeast begins to recycle the alcohol it previously produced into stored sugars. By retaining some ancestral function and losing redundant function over millions of years of evolution, *TDH1* and *TDH2* would be maintained in the *S. cerevisiae*

genome because they could still be used for carrying out enzymatic reactions when they are needed, or for some other unknown function.

Another explanation for what could be going on with the differential expression of *TDH* paralogs is that there could be differences in how the genes' protein products interact with each other. TDH3 has been known to form a heteromer with TDH1 and TDH2 (Randez-Gil et al. 2019) to carry out cellular functions. What these cellular functions are is unclear. In work done by Marchant et al. (2019), it was found that the protein products of paralogs with high sequence similarity tend to form heteromers. This has been proposed with the protein products of the *TDH*s and it is important to consider when thinking about why and how the *TDH*s function in the way that they do. If the observed subfunctionalization of *TDH* regulatory sequences following gene duplication is because all three TDH proteins are needed for heteromer formation then selection would favor the differential expression of the paralogs.

Some other constraints must be considered when interpreting the data that was collected, and most were due to time constraints and failed experiments. Due to the long process of designing a CRISPR-Cas9 experiment in yeast, there were times during this thesis research when weeks were dedicated to designing and obtaining the small pieces needed to make sure that experiments were efficient (gRNA design, repair fragment design, primer design, etc.). With these things also came failed experiments. Not every CRISPR-Cas9 engineering attempt worked on the first try; in fact, there may have been many failed multiple attempts before a successful transfectant was obtained. This led to better optimizing experiments, which contributed to a lag in the time spent conducting the wet-lab procedures necessary for strain generation. Therefore, in the end, if the field of evolutionary genetics wants to continue the work in understanding how and why the *TDH* paralogs

in yeast are differentially expressed yet highly conserved, more time and energy must be spent on phenotyping and characterizing each paralogs' contribution to yeast cell fitness.

Potential novel functions of *TDH* protein products following millions of years of evolution in *S. cerevisiae*

Introduction

Gene duplication events may give rise to new, previously unknown functions through the process of neofunctionalization. Examples of this include the diversification of plants that undergo C_4 photosynthesis (Monson 2003), the evolution of snake venom genes (Casewell et al. 2011), and retinoic acid receptors in vertebrates (Escriva et al. 2006). Neofunctionalization is fascinating and important in its own right, but it can also take place in tandem with or following subfunctionalization (He & Zhang 2005). The *Saccharomyces cerevisiae* genome, which is likely the result of whole-genome duplication, is a great example of how this might be happening.

TDH1, TDH2, and *TDH3* are paralogs of an ancestral *TDH* gene. These paralogs encode the enzyme glyceraldehyde-3-phosphate dehydrogenase, which catalyzes the sixth step of glycolysis and is essential for yeast survival. Interestingly, these genes are both highly conserved and have large sequence similarities, yet they are differentially expressed. An explanation for this could be that due to functional redundancy, they are expressed at varying levels so that duplicates are maintained. A form of subfunctionalization, called expression reduction, speaks to this (Qian et al 2010). This model of subfunctionalization works to explain why functionally redundant genes have been kept following millions of years of evolution, and this might explain why all of the *TDH* paralogs have been maintained in *S. cerevisiae*.

If *TDH* genes are not being differentially expressed due to expression reduction, the paralogs may have been maintained due to neofunctionalization. In Chapter 1, it was shown that the *TDH3* promoter, P_{TDH3}, evolved to cause disproportionate contributions of the TDH3 protein to fitness in *S. cerevisiae*. If the changes in this promoter caused it to be more highly expressed

because P_{TDH2} and P_{TDH1} are subject to expression reduction, the differences in expression levels would be explained. However, if regulatory changes following gene duplication caused novel protein functions to arise, neofunctionalization of *TDH* protein-coding sequences might have to be considered along with expression reduction.

In this chapter, I seek to better understand the consequences of the regulatory changes that arose in *TDH* paralogs following gene duplication. It is known that these paralogs are differentially expressed, and we now have a better understanding as to why that is, and I wanted to ask the question as to how regulatory changes might affect other phenotypes beyond exponential growth. Specifically, I wanted to investigate whether phenotypes distinct from glycolysis, which happens in the cytoplasm, may have contributed to functional divergence of the TDH proteins. To test this, I first tagged both wildtype *TDH* paralogs and chimeric promoter/protein-coding sequence swap strains with *ScarletI*, a fluorophore, to test for nuclear inclusion or exclusion. I also tagged *TDH1, TDH2*, and *TDH3* with fluorophores with distinct and separate emission spectra in the same cell, therefore creating a triple fluorophore yeast strain. Following this, I fixed and imaged the yeast with a Leica SP8 confocal microscope to see if the paralogs' protein products localized to different areas of the cell.

Materials and Methods

Much of this section would not have been possible without the help of Hannah Kania. She was instrumental in the creation of most TDH-ScarletI *yeast strains, as she designed many of the gRNAs and repair fragments needed for the CRISPR-Cas9 engineering of the protein-fluorophore chimeras. Hannah is now a Ph.D. student at Duke University studying evolutionary genetics.*

***Construction of* TDH-ScarletI *chimeric yeast strains*.** To create the *TDH-ScarletI* chimeric strains of *S. cerevisiae*, a CRISPR-Cas9 genome editing protocol adapted from Laughery et al. 2015 was used. Wild-type *S. cerevisiae* + Δ*URA3* was grown up in 5 mL of YPD overnight at

30℃, and diluted to an OD of 0.75 before the experiments were conducted. CRISPR-Cas9 yeast transfections were done in which gRNAs targeted to the C-terminus of *TDH* protein-coding sequences were used with Cas9 to cut the C-termini and fuse *ScarletI* onto them (Figure 2.1). gRNAs were stored in pML104 plasmids that also included *URA3* so as to allow transfected colony growth on selective media. Following CRISPR-Cas9 transfections, transfected yeast were plated onto SC-URA agar plates to select for successful transfectants. Plates were incubated for two days. Following this period, the plates were observed for successful transfectants. Successful transfectants fluoresce red under an RFP filter, indicating that the ScarletI fluorophore fused onto the C-terminus of the *TDH* gene that was tagged. Observed colonies were moved to 5'-FOA plates to remove residual genome-editing plasmids from the colonies by selecting for the absence of *URA3*. These plates were incubated for 2 days at 30°C. Following the incubation period, colonies were checked for the presence of the correct genotype by PCR amplification and gel electrophoresis. Successful transfectants were then sent to Eurofins Genomics for sequencing. With the confirmation of successful transfection, new yeast strains were glycerol stocked in 20% glycerol and placed into a -80℃ freezer for storage. Strains are represented in Table 1 (Appendix).

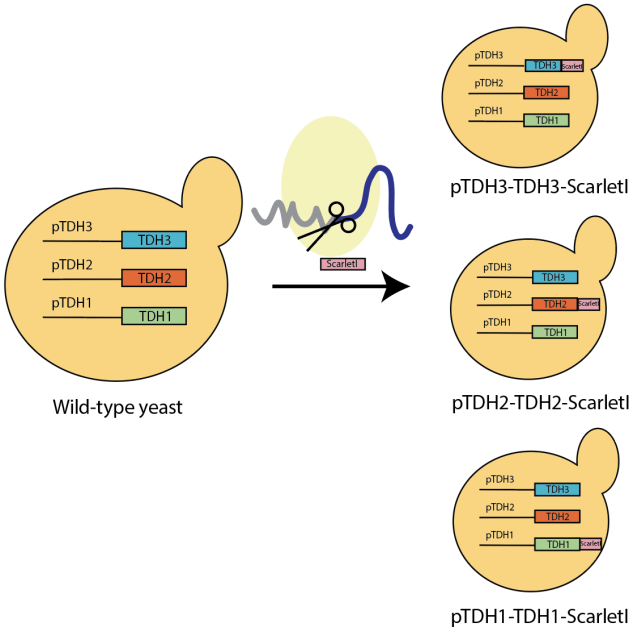

Figure 2.1: Creation of chimeric *TDH-ScarletI* strains of *S. cerevisiae*. *TDH* genes were tagged with *ScarletI* to test for the localization of *TDH* protein products. *ScarletI* repair fragments were made using PCR amplification of primers with 5' and 3' homology to the *TDH* being targeted. These repair fragments were then utilized in the CRISPR-Cas9 experiment to integrate them onto the C-termini of *TDH*s. *TDH* chimeric strains were sequence verified.

***Construction of triple chimeric protein-fluorophore yeast strain*.** The triple chimeric protein-fluorophore yeast strain was made in the background of yeast strain 3972 (Table 1, Appendix). This strain already had *ScarletI* tagged onto the C-terminus of *TDH2*, so all that needed to be done was tagging of *YFP* to *TDH1* and *Tq2* to *TDH3*. These fluorophores were chosen intentionally; the weakest fluorophore (*Tq2*) was tagged to the highest expressing gene (*TDH3*), and the strongest fluorophore (*YFP*) was tagged to the weakest expressing gene (*TDH1*) to account for discrepancies in fluorophore feedback. To tag *Tq2* to *TDH3*, first PCR was used to amplify fluorophore out of a plasmid created by Botman et al. (2019), and the primers that were used to amplify the fluorophore

had 5' and 3' homology arms to the C-terminus of *TDH3*. By using these primers, *Tq2* was amplified out of the plasmid and the amplicon had homology to *TDH3*. To insert the fluorophore amplicon into yeast strain 3972's *TDH3* locus, a CRISPR-Cas9 experiment was performed. Following the experiment, the colonies were PCR screened and sequence verified. The resulting strain was numbered 3983 and it was stored in a 20% glycerol solution in a -80°C freezer.

To tag *YFP* to *TDH1*, *YFP* was amplified out of yeast strain 3293 (Table 1, Appendix), which had *TDH1* fused to *YFP*. By amplifying the *TDH1-YFP* sequence out of 3293, homology to the C-terminus of *TDH1* in yeast strain 3983 was ensured. Once *TDH1-YFP* was amplified, I used CRISPR-Cas9 to delete *TDH1* out of 3983, and then I inserted the *TDH1-YFP* amplicon at the same locus. This strain was sequence verified, given the number 3984, and stored in a 20% glycerol solution in a -80°C freezer.

***Fixing and DAPI-staining of* TDH-ScarletI *fusion yeast strains*.** To prepare cells for imaging, they were first grown in 5 ml YPD overnight at 30°C. The following morning, cells were centrifuged for 2 minutes at 5000 rpm. The supernatant was discarded and cells were resuspended in 1 mL PBS. Cells were then centrifuged again for 2 minutes at 5000 rpm. The supernatant was discarded and cells were resuspended in .250 mL 4% paraformaldehyde in PBS. Cells were moved to a benchtop shaker, covered with foil, and left to shake for 20 minutes at room temperature. Following the shaking period, cells were centrifuged for 2 minutes at 5000 rpm. The supernatant was discarded and cells were resuspended in a 1% PBST solution and left to incubate covered in foil on a benchtop at room temperature. Following the incubation, cells were centrifuged at 5000 rpm for 2 minutes. The supernatant was removed, and cells were resuspended in 0.5 mg/mL DAPI in PBST and left to shake covered in foil on a benchtop shaker at room temperature for 30 minutes.

Following shaking, cells were centrifuged for 2 minutes at 5000 rpm. The supernatant was discarded and cells were resuspended in .250 mL PBS. Cells were then stored in a fridge at 4°C for up to 3 days or until cells were imaged.

Fixing of triple chimeric protein-fluorophore yeast strain. To prepare cells for imaging, they were first grown in 5 ml YPD overnight at 30°C. The following morning, cells were centrifuged for 2 minutes at 5000 rpm. The supernatant was discarded and cells were resuspended in 1 mL PBS. Cells were then centrifuged again for 2 minutes at 5000 rpm. The supernatant was discarded and cells were resuspended in .250 mL 4% paraformaldehyde in PBS. Cells were moved to a benchtop shaker, covered with foil, and left to shake for 20 minutes at room temperature. Following the shaking period, cells were centrifuged for 2 minutes at 5000 rpm. The supernatant was discarded and cells were resuspended in a 50% glycerol solution in PBS and left to incubate covered in foil on a benchtop at room temperature. Following the incubation, cells were centrifuged at 5000 rpm for 2 minutes. The supernatant was removed, and cells were resuspended in a 50% glycerol solution in PBS. Cells were then stored in a fridge at 4°C for up to 3 days or until cells were imaged.

Results

TDH proteins localize to the nucleus when protein-coding sequences are driven by $\mathbf{P_{TDH3}}$

The nucleus is worth looking at in yeast because if TDH proteins are seen to be localizing there, the proteins may have some unknown nuclear function. To determine whether differences in expression of each *TDH* paralog contributed to novel nuclear functions of TDH proteins, a strain of *S.*

cerevisiae was constructed in which a red fluorescent protein (RFP) called ScarletI was fused to the end of the paralogs' protein-coding sequence. Following the fusion of ScarletI with CRISPR-Cas9, chimeric yeast strains were grown in 5 mL YPD for twenty-four hours. Following growth, strains were washed with PBS[2], fixed with 4% paraformaldehyde (PFA)[3], and stained with DAPI[4] for nuclear visualization. All yeast strains that were visualized had yellow fluorescent protein, or YFP, constitutively expressed by the *HO* locus.[5] When ScarletI was fused to the end of *TDH3* and the protein-fluorophore chimera was visualized, it was apparent that TDH3 was traveling to the nucleus (Figure 2.2). Similarly, when TDH2 is visualized following the construction of a *TDH2-ScarletI* chimeric strain, TDH2 can be found in the nucleus (Figure 2.3). In contrast, when *TDH1* is fused to *ScarletI* and the chimeric protein is visualized, it is not found in the nucleus (Figure 2.4).

I next tested whether expressing other TDH proteins at the level of TDH3 is sufficient for them to be localized to the nucleus. To test whether expression under the control of P_{TDH3} causes localization to the nucleus, chimeric P_{TDH3}-*TDH-ScarletI* strains were constructed with CRISPR-Cas9 genome editing technology. These strains were fixed and stained using the same protocol that was used for the visualization of individual *TDH* genes. When *ScarletI* was fused to the end of P_{TDH3}-*TDH2* and the protein product was visualized, TDH2 was found in the nucleus (Figure 2.5). No visual data is available for the P_{TDH3}-*TDH1-ScarletI* strain because it was not engineered

[2] PBS is a saline buffer that washes cells of debris and ensures that they do not rupture due to the osmotic pressure of the solution they might be in.
[3] PFA fixes cells by cross-linking proteins in the cells.
[4] DAPI is a blue fluorescent stain that binds to AT regions of DNA. It is a good fluorophore to use for visualization of nuclear inclusion or exclusion of other fluorescent proteins due to the fact that it is easily visible when nuclei are being visualized by microscopy.
[5] The *HO* locus in yeast is the mating-type locus. By using the *HO* locus to drive YFP expression, YFP is always produced and should be everywhere in the yeast cells.

at the time of visualization. P_{TDH3}-*TDH2-ScarletI* does not appear to be visually different from P_{TDH2}-*TDH2-ScarletI*.

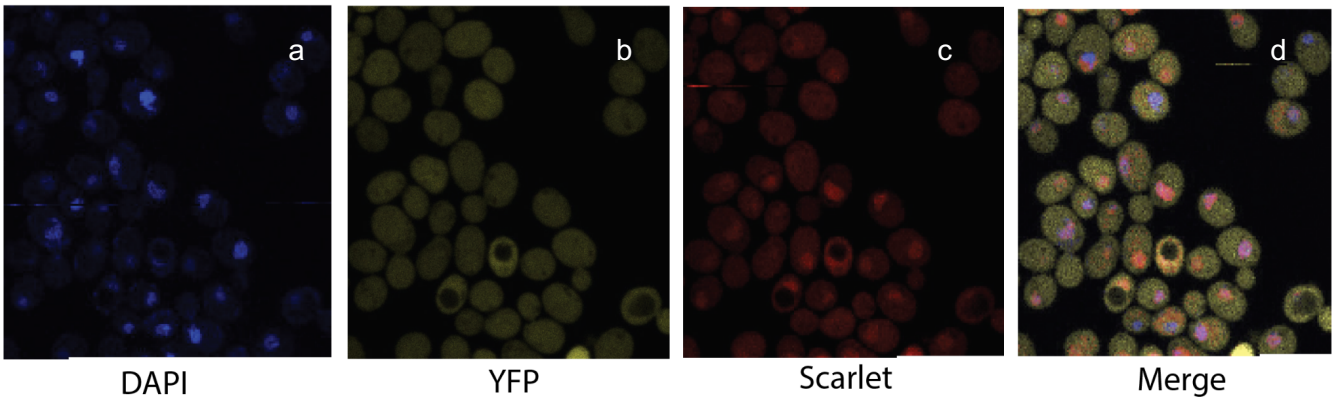

Figure 2.2: **Localization of TDH3 to the nucleus.** Wild-type TDH3 localizes to the nucleus, as shown by *TDH3-Scarlet* fusion. (a) DAPI nuclear stain of cells. (b) YFP is constitutively expressed by the *HO* locus in *S. cerevisiae*. (c) TDH3-Scarlet fusion protein is present in the cytoplasm and the nucleus of cells. (d) Merged overlay of all fluorophores.

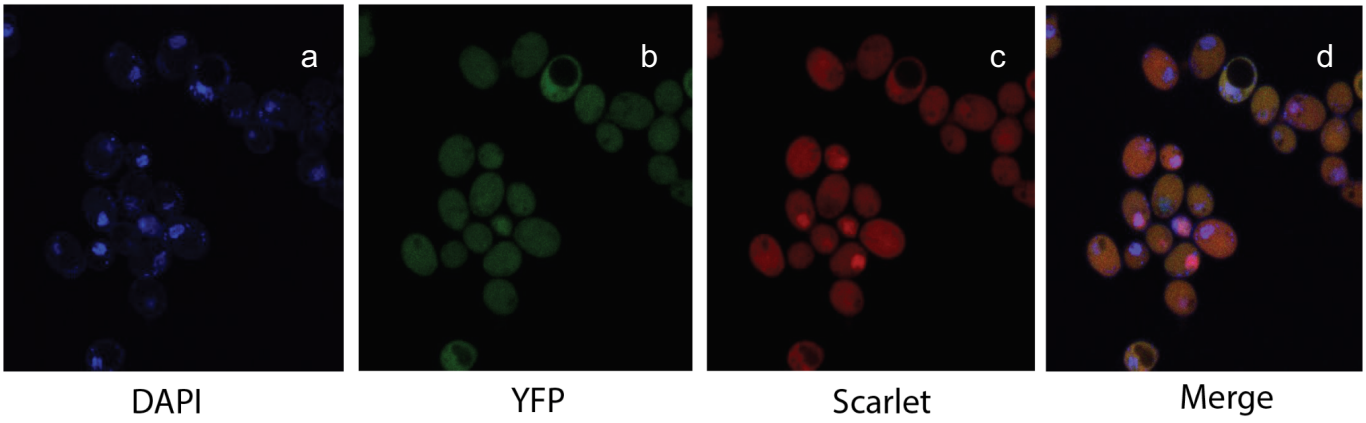

Figure 2.3: **Localization of TDH2 to the nucleus.** Wild-type TDH2 localizes to the nucleus, as shown by *TDH2-Scarlet* fusion.(a) DAPI nuclear stain of cells. (b) YFP is constitutively expressed by the *HO* locus in *S. cerevisiae*. (c) TDH2-Scarlet fusion protein is present in the cytoplasm and the nucleus of cells. (d) Merged overlay of all fluorophores.

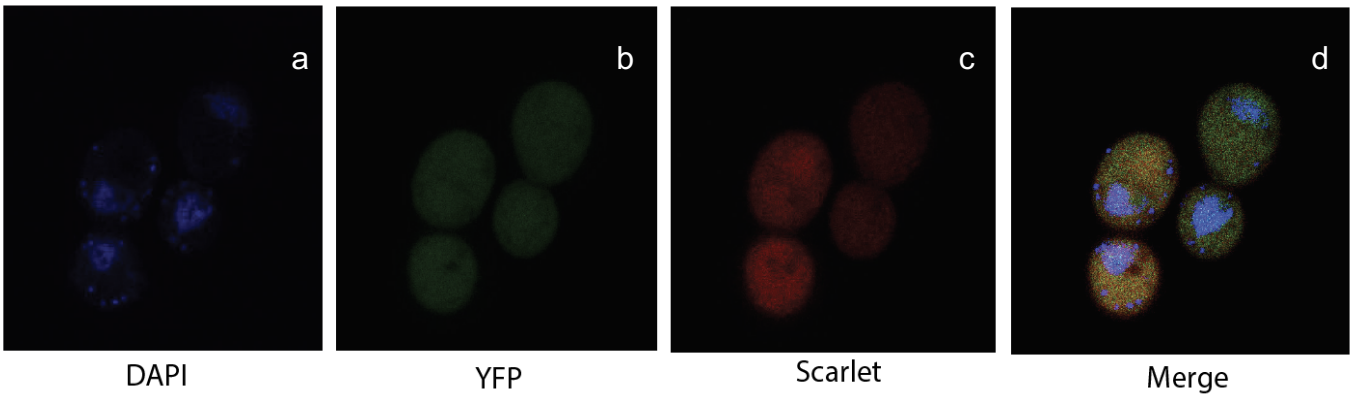

Figure 2.4: **TDH1 does not localize to the nucleus.** Wild-type TDH1 does not localize to the nucleus, as shown by *TDH1-Scarlet* fusion. (a) DAPI nuclear stain of cells. (b) YFP is constitutively expressed by the *HO* locus in *S. cerevisiae*. (c) TDH1-Scarlet fusion protein is present only in the cytoplasm cells. (d) Merged overlay of all fluorophores.

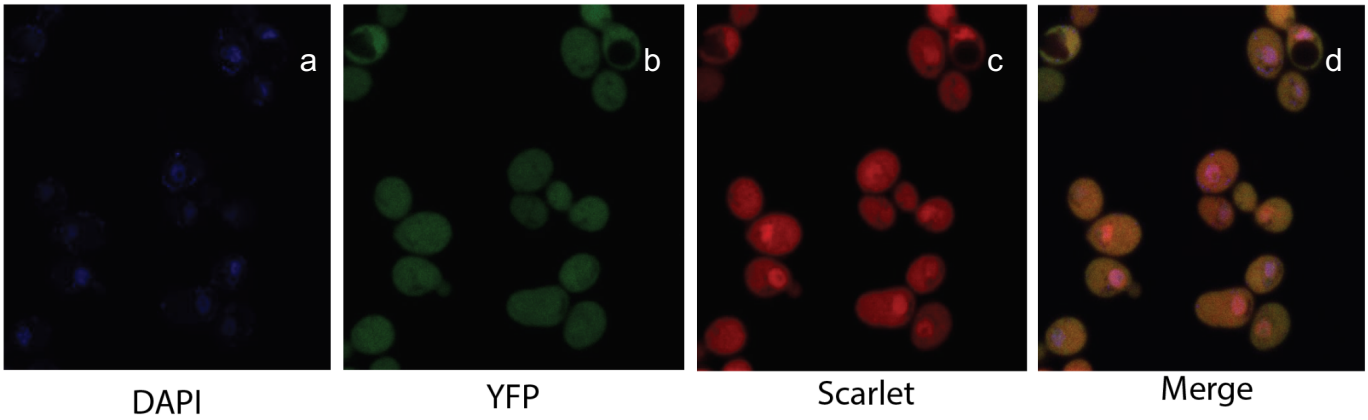

Figure 2.5: **Localization of TDH2 to the nucleus when driven by P_{TDH3}.** TDH2 localizes to the nucleus when driven by P_{TDH3} at the native *TDH3* locus. (a) DAPI nuclear stain of cells. (b) YFP is constitutively expressed by the *HO* locus in *S. cerevisiae*. (c) TDH2-Scarlet fusion protein is present in the cytoplasm and the nucleus of cells when expression is driven by P_{TDH3}. (d) Merged overlay of all fluorophores.

Triple fluorophore yeast strain reveals that all three TDH proteins localize to the nucleus after 48 hours of growth

To further investigate differences in TDH localization, a chimeric strain of yeast was engineered in which each *TDH* paralog was tagged with a different fluorophore with distinct and separate emission spectra. *TDH1* was fused to *YFP*, *TDH2* was fused to *ScarletI*, and *TDH3* was fused to *Turquoise2* (*Tq2*).[6] All of these fusions were done by individually engineering them with CRISPR-Cas9. In contrast to the proteins that were visualized individually with ScarletI, YFP was not constitutively expressed by the *HO* mating-type locus because any background YFP would interfere with the signal being produced by *TDH1* expression. Also, a DAPI nuclear stain was not done because it would interfere with the fluorophore emissions of *TDH3* expression.

Following the construction of the triple chimeric protein-fluorophore strain, the strain was grown up in 5 mL of YPD for forty-eight hours. Following growth, the yeast were washed with PBS and fixed with 4% paraformaldehyde. Visualization of the strain was done using confocal microscopy on a Leica SP8 scope. Imaging revealed that after forty-eight hours of cell culture growth, each *TDH* paralog protein product can be found in the nucleus (Figure 2.6).

[6] *Tq2* encodes a cyan fluorescent protein, or CFP, that fluoresces light blue when excited. It was made by Botman et al. (2019)

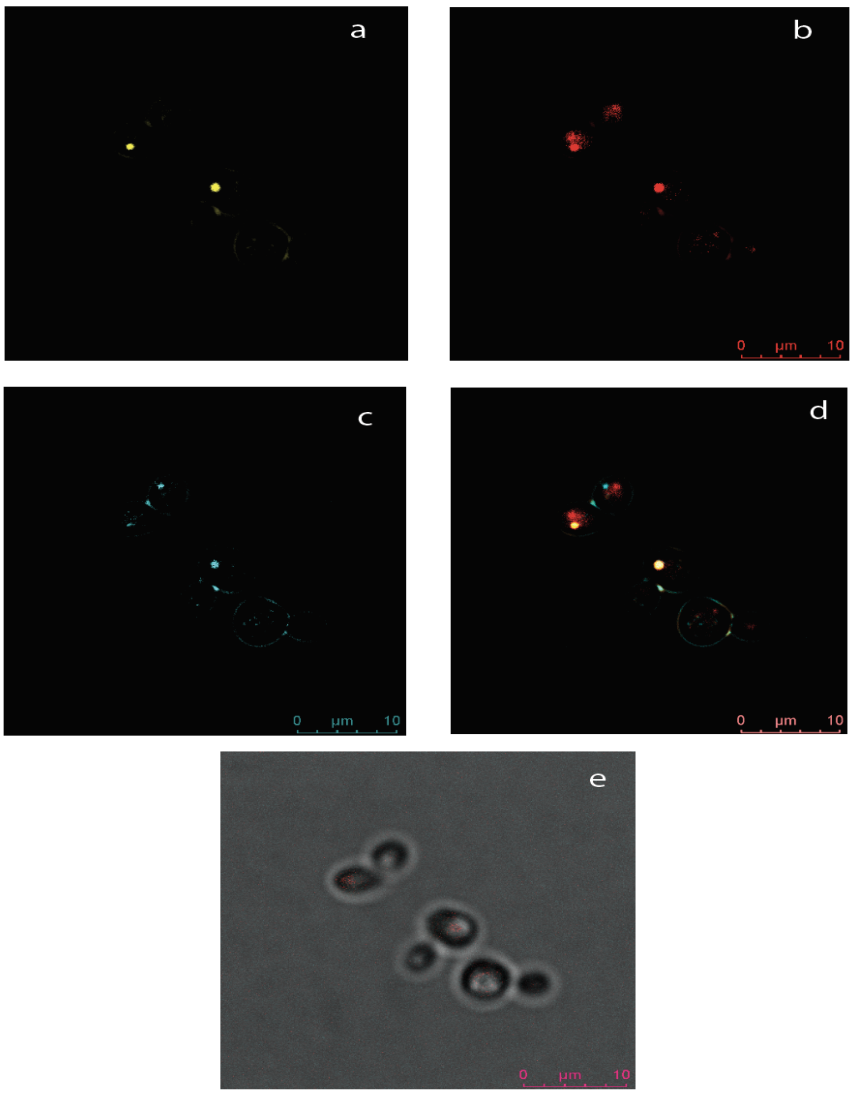

Figure 2.6: *TDH* protein products localize to the nucleus following 48 hours of growth. (a) YFP fused to TDH1 shows nuclear localization. (b) ScarletI fused to TDH2 shows both nuclear and cytoplasmic localization. (c) Tq2 fused to TDH3 shows nuclear localization. (d) Merged overlay of fluorophore localization shows nuclear localization of all TDH proteins. (e) Brightfield image of triple fluorophore yeast.

Discussion

Subfunctionalization of gene duplicates has been proposed to be a transition state to neofunctionalization (Rastogi & Liberles 2005). As was previously explored, glycolytic gene paralogs *TDH1, TDH2,* and *TDH3* in *S. cerevisiae* show evidence that all three paralogs are maintained through the subfunctionalization of the regulatory sequences of the genes. TDH3 is the most active protein product in yeast cells with 50 to 60% of glycolytic activity being attributed to it (McAlister and Holland 1985), with TDH2 and TDH1 falling behind it. The down-regulation of *TDH1* and *TDH2* and therefore the lack of abundance and activity of these proteins could explain how subfunctionalization has functioned to maintain the gene copies; however, one must also consider that the proteins have evolved novel functions following the hundreds of millions of years following the gene duplication event that gave rise to the paralogs. In this chapter of this thesis, I attempted to characterize how possible neofunctionalization of *TDH* protein-coding sequences may be happening.

New protein functions may evolve that are cellularly distinct from glycolysis. If so, the proteins would be localized elsewhere. To test whether the differences in expression of *TDH* paralogs are due to the fact that the protein products are functionally different peptides, I tagged the protein-coding sequence of each *TDH* gene with fluorophores to test and see if nuclear inclusion or exclusion was taking place in the cells. Glycolysis takes place in the cytoplasm, so TDH proteins would likely be there if they were still carrying out the ancestral function of the conversion of glyceraldehyde 3-phosphate to 1,3-bisphosphoglycerate. Interestingly, that is not what I saw. After twenty-four hours of growth, I saw that TDH3 and TDH2, but not TDH1, are found almost exclusively in the nucleus. TDH3 was found in the nucleus more often than TDH2, but it was still there. Along with this, when the *TDH3* promoter is driving the expression of *TDH2*

at the native *TDH3* locus, TDH2 is found in the nucleus. Furthermore, I found that following forty-eight hours of yeast growth, the protein products of all three paralogs can be found in the nucleus. This is curious, as the protein products would be expected to be in the cytoplasm carrying out their role in glycolysis. There are a few reasons as to why *TDH1*, *TDH2*, and *TDH3* may have diverged in some ways from their ancestral function, but before discussing them I must take into consideration the limitations of the data.

The first thing that needs to be considered is the fact that due to time limitations and experimental constraints, not every *TDH* promoter/protein-coding sequence chimera that was made for phenotyping observed in the first chapter was able to be engineered for visualization. This lack of visual data makes it challenging to draw significant conclusions about what is going on with the localization of *TDH* protein products. Also, yeast strains in which the fluorophores are expressed by the native *TDH* promoters without the presence of the protein-coding sequences were not made. This makes it difficult to tell if the TDH proteins are truly traveling to the nucleus, or if they are only traveling there because they were tagged with a fluorophore. The fluorophores were also not attached to a protein that is known to be purely cytoplasmic, so neglecting to do this also makes it hard to conclude the nuclear inclusion or exclusion of *TDH* protein products. Another limitation of these experiments is that the field of evolutionary genetics does not know the exact function of each *TDH* protein product, therefore making it difficult to make complete sense of the visual data that was obtained. While my data shows that there is potential divergence in function between the paralogous proteins, that function is not known and more experiments need to be done to elucidate exact functions. Such experiments might include chromatin immunoprecipitation following sequencing, or ChIP-sequencing, which would determine if there are TDH protein interactions with open chromatin in the nucleus. My visual data shows that TDH proteins are

localizing to the nucleus, and ChIP-sequencing would be a good next step in determining why that might be (Ringel et al. 2013). Limitations aside, there are some explanations for why *TDH* protein products are localized to the nucleus, and nuclear activities of GAPDH proteins have been reported in studies from humans and other organisms (Sirover 2005).

Following gene duplication, subfunctionalization can lead to the asymmetric evolution of paralogs' protein products in *S. cerevisiae* (Byrne & Wolfe 2007). Whether this asymmetric evolution has caused proteins to take on novel functions is unclear, but there are cases in which TDH proteins have been noted to participate in cellular processes not even remotely tied to glycolysis in microbes. One such observation is TDH protein activity as an antimicrobial peptide (AMP) (Pereira et al. 2021, Branco et al. 2013). When grown in a sterilized grape juice culture with another yeast strain, *Hanseniaspora guilliermondii*, *S. cerevisiae*'s TDH proteins act to kill off *H. guilliermondii* in the culture (Branco et al. 2013). This was confirmed by *TDH* protein-coding sequence deletions; when *TDH2* and *TDH3* were deleted in *S. cerevisiae*, it was less capable of killing off *H. guilliermondii* in a mixed culture. This is especially interesting because *TDH3* and *TDH2* are the two genes most essential for *S. cerevisiae* viability during rapid growth. My growth phenotyping data shows that the *TDH* paralogs potentially underwent regulatory subfunctionalization following gene duplication, and with this, my microscopy data shows that the protein-coding sequences of the paralogs may have undergone neofunctionalization following regulatory subfunctionalization. This hypothesis has been backed up by the Branco group (2013) and the Pereira group (2021), and other researchers have also proposed that TDH proteins have more than just ancestral glycolytic function (Nakajima et al. 2009, Silva et al. 2011).

The data presented in the chapter, while important, is inconclusive. Because the exact functions of each *TDH* paralog are unknown, any visual data that was collected is purely

speculative and will continue to be until further research is done. However, the data presented does bring up an important question that is highly relevant to the field of evolutionary genetics: Did glycolytic gene paralogs *TDH1*, *TDH2*, and *TDH3* undergo the neofunctionalization of their protein-coding sequences following the subfunctionalization of their regulatory sequences, and if so, what are the novel functions that arose over hundreds of millions of years of evolution? To fully answer this question more research needs to be done, but the work presented in this chapter lays the groundwork for the research to be successfully continued.

Conclusion and Future Directions

The research presented in this thesis focuses on why the glycolytic gene paralogs *TDH1*, *TDH2*, and *TDH3* have been maintained in *Saccharomyces cerevisiae*. I investigated whether this was due to changes in their regulatory sequences or through changes in their protein-coding sequences. I engineered a multitude of *S. cerevisiae* strains with *TDH* protein-coding sequences deleted and with chimeric promoter/protein-coding sequence swaps to test this question, and I found that *TDH* genes are differentially expressed due to potential subfunctionalization of their regulatory sequences. Furthermore, I investigated whether or not the subfunctionalization of the regulatory regions was followed by the neofunctionalization of the protein-coding sequences by visualizing TDH protein products tagged to fluorophores. Here, I found that TDH2 and TDH3 localize to the nucleus following twenty-four hours of cell growth and that all three protein products localize to the nucleus following forty-eight hours of cell growth. This raises the tantalizing possibility that new functions unrelated to glycolysis may be evolving and diversifying in the paralogs. The mechanisms through which this is happening are unknown, but I proposed experiments and logic as to why it may be occurring.

This thesis attempts to unravel some of the questions surrounding gene duplication that the field of evolutionary genetics has been attempting to answer for years. While some of the data yields promising answers, many more questions arose following the research presented here. In the future, experiments should be conducted in which more phenotyping of *TDH* deletions and promoter/protein-coding sequence swaps is done so that a more clear picture of the subfunctionalization of the *TDH* paralogs can be drawn. Along with this, work to visualize more *TDH* protein products should be done. This could include visualizing the proteins on a time-lapse to better understand where the proteins are localizing when, and it might also include

characterizing TDH protein-DNA interactions through ChIP-sequencing of the *S. cerevisiae* genome to see if there are such interactions occurring.

Overall, much is to be uncovered about the mechanisms through which gene duplicates are kept in *S. cerevisiae* and whether such mechanisms occur in tandem with each other or individually following selection. Further work, including the experiments described above, will help to elucidate information about the *TDH* paralogs in yeast and the ways in which they have stayed conserved in the genome following hundreds of millions of years of selection.

References

Barkman, T., & Zhang, J. (2009). Evidence for escape from adaptive conflict? *Nature*, *462*(7274). https://doi.org/10.1038/nature08663

Boonekamp, F. J., Dashko, S., van den Broek, M., Gehrmann, T., Daran, J.-M., & Daran-Lapujade, P. (2018). The genetic makeup and expression of the glycolytic and fermentative pathways are highly conserved within the saccharomyces genus. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/fgene.2018.00504

Botman, D., de Groot, D. H., Schmidt, P., Goedhart, J., & Teusink, B. (2019). In vivo characterisation of fluorescent proteins in budding yeast. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-38913-z

Branco, P., Francisco, D., Chambon, C., Hébraud, M., Arneborg, N., Almeida, M. G., Caldeira, J., & Albergaria, H. (2013). Identification of novel GAPDH-derived antimicrobial peptides secreted by saccharomyces cerevisiae and involved in wine microbial interactions. *Applied Microbiology and Biotechnology*, *98*(2), 843–853. https://doi.org/10.1007/s00253-013-5411-y

Byrne, K. P., & Wolfe, K. H. (2007). Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, *175*(3), 1341–1350. https://doi.org/10.1534/genetics.106.066951

Casewell, N. R., Wagstaff, S. C., Harrison, R. A., Renjifo, C., & Wuster, W. (2011). Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. *Molecular Biology and Evolution*, *28*(9), 2637–2649. https://doi.org/10.1093/molbev/msr091

Crow, K. D. (2006). What is the role of genome duplication in the evolution of complexity and Diversity? *Molecular Biology and Evolution*, *23*(5), 887–892. https://doi.org/10.1093/molbev/msj083

Doudna, J. A., & Charpentier, E. (2014). The New Frontier of Genome Engineering with CRISPR-Cas9. *Science*, *346*(6213). https://doi.org/10.1126/science.1258096

Gonçalves, P., & Planta, R. J. (1998). Starting up yeast glycolysis. *Trends in Microbiology*, *6*(8), 314–319. https://doi.org/10.1016/s0966-842x(98)01305-5

Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity*, *100*(5), 605–617. https://doi.org/10.1093/jhered/esp047

Hahn, M. W. (2009). Distinguishing among evolutionary models for the maintenance of gene duplicates. *Journal of Heredity*, *100*(5), 605–617. https://doi.org/10.1093/jhered/esp047

He, X., & Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, *169*(2), 1157–1164. https://doi.org/10.1534/genetics.104.037051

Lan, X., & Pritchard, J. K. (2016). Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science*, *352*(6288), 1009–1013. https://doi.org/10.1126/science.aad8411

Laughery, M. F., Hunter, T., Brown, A., Hoopes, J., Ostbye, T., Shumaker, T., & Wyrick, J. J. (2015). New vectors for simple and streamlined CRISPR-Cas9 genome editing in *Saccharomyces cerevisiae*. *Yeast*, *32*(12), 711–720. https://doi.org/10.1002/yea.3098

Lobo, I. (2008) Pleiotropy: One Gene Can Affect Multiple Traits. *Nature Education* 1(1):10

Lynch, M. (2002). Gene duplication and evolution. *Science*, *297*(5583), 945–947. https://doi.org/10.1126/science.1075472

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, *154*(1), 459–473. https://doi.org/10.1093/genetics/154.1.459

Marchant, A., Cisneros, A. F., Dubé, A. K., Gagnon-Arsenault, I., Ascencio, D., Jain, H., Aubé, S., Eberlein, C., Evans-Yamamoto, D., Yachie, N., & Landry, C. R. (2019). The role of structural pleiotropy and regulatory evolution in the retention of heteromers of paralogs. *ELife*, *8*. https://doi.org/10.7554/elife.46754

McAlister, L., & Holland, M. J. (1985). Differential expression of the three yeast glyceraldehyde-3-phosphate dehydrogenase genes. *Journal of Biological Chemistry*, *260*(28), 15019–15027. https://doi.org/10.1016/s0021-9258(18)95696-6

Monson, R. K. (2003). Gene duplication, Neofunctionalization, and the evolution of C4 Photosynthesis. *International Journal of Plant Sciences*, *164*(S3). https://doi.org/10.1086/368400

Nakajima, H., Amano, W., Kubo, T., Fukuhara, A., Ihara, H., Azuma, Y.-T., Tajima, H., Inui, T., Sawa, A., & Takeuchi, T. (2009). Glyceraldehyde-3-phosphate dehydrogenase aggregate formation participates in oxidative stress-induced cell death. *Journal of Biological Chemistry*, *284*(49), 34331–34341. https://doi.org/10.1074/jbc.m109.027698

Ohno, S. (1970). *Evolution by gene duplication*. Springer.

Pereira, P. R., Freitas, C. S., & Paschoalin, V. M. (2021). *saccharomyces cerevisiae* biomass as a source of next-generation food preservatives: Evaluating potential proteins as a source of antimicrobial peptides. *Comprehensive Reviews in Food Science and Food Safety*, *20*(5), 4450–4479. https://doi.org/10.1111/1541-4337.12798

Qian, W., Liao, B.-Y., Chang, A. Y.-F., & Zhang, J. (2010). Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends in Genetics*, *26*(10), 425–430. https://doi.org/10.1016/j.tig.2010.07.002

Randez-Gil, F., Sánchez-Adriá, I. E., Estruch, F., & Prieto, J. A. (2019). The formation of hybrid complexes between isoenzymes of glyceraldehyde-3-phosphate dehydrogenase regulates its aggregation state, the glycolytic activity and

sphingolipid status in *Saccharomyces cerevisiae*. *Microbial Biotechnology*, *13*(2), 562–571. https://doi.org/10.1111/1751-7915.13513

Rastogi, S., & Liberles, D. A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology*, *5*(1). https://doi.org/10.1186/1471-2148-5-28

Ringel, A. E., Ryznar, R., Picariello, H., Huang, K.-lin, Lazarus, A. G., & Holmes, S. G. (2013). Yeast TDH3 (glyceraldehyde 3-phosphate dehydrogenase) is a SIR2-interacting factor that regulates transcriptional silencing and rdna recombination. *PLoS Genetics*, *9*(10). https://doi.org/10.1371/journal.pgen.1003871

Silva, A., Almeida, B., Sampaio-Marques, B., Reis, M. I. R., Ohlmeier, S., Rodrigues, F., Vale, A. do, & Ludovico, P. (2011). Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is a specific substrate of yeast metacaspase. *Biochimica Et Biophysica Acta (BBA) - Molecular Cell Research*, *1813*(12), 2044–2049. https://doi.org/10.1016/j.bbamcr.2011.09.010

Sirover, M. A. (2005). New nuclear functions of the glycolytic protein, glyceraldehyde-3-phosphate dehydrogenase, in mammalian cells. *Journal of Cellular Biochemistry*, *95*(1), 45–52. https://doi.org/10.1002/jcb.20399

van Hoof, A. (2005). Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication. *Genetics*, *171*(4), 1455–1461. https://doi.org/10.1534/genetics.105.044057

Voordeckers, K., Pougach, K., & Verstrepen, K. J. (2015). How do regulatory networks evolve and expand throughout evolution? *Current Opinion in Biotechnology*, *34*, 180–188. https://doi.org/10.1016/j.copbio.2015.02.001

Zhang, J. (2003). Evolution by gene duplication: An update. *Trends in Ecology & Evolution*, *18*(6), 292–298. https://doi.org/10.1016/s0169-5347(03)00033-8

Appendix

Table 1: Generated *S. cerevisiae* strains used for thesis research

Strain number	TDH1	TDH2	TDH3	Species	Project
3293	pTDH1-TDH1-YFP	pTDH2-TDH2	pTDH3-TDH3	*S. cerevisiae*	Triple fluorophore
3324	pTDH1-TDH1	pTDH2-TDH2	pTDH3-TDH3	*S. cerevisiae*	All
3821	pTDH1-TDH1	pTDH2-ΔTDH2	pTDH3-TDH3	*S. cerevisiae*	Protein Deletions/Swaps
3822	pTDH1-ΔTDH1	pTDH2-TDH2	pTDH3-TDH3	*S. cerevisiae*	Protein Deletions/Swaps
3824	pTDH1-TDH1	pTDH2-TDH2	pTDH3-ΔTDH3	*S. cerevisiae*	Protein Deletions/Swaps
3825	pTDH1-TDH1	pTDH2-TDH2	pTDH3-TDH2	*S. cerevisiae*	Protein Deletions/Swaps
3838	pTDH1-TDH1	pTDH2-ΔTDH2	pTDH3-ΔTDH3	*S. cerevisiae*	Protein Deletions/Swaps
3840	pTDH1-TDH1	pTDH2-ΔTDH2	pTDH3-TDH2	*S. cerevisiae*	Protein Deletions/Swaps
3841	pTDH1-TDH1	pTDH2-TDH2	pTDH3-TDH2-Scarlet	*S. cerevisiae*	Scarlet Localization
3842	pTDH1-TDH1	pTDH2-TDH2	pTDH3-TDH1	*S. cerevisiae*	Protein Deletions/Swaps
3843	pTDH1-ΔTDH1	pTDH2-ΔTDH2	pTDH3-TDH3	*S. cerevisiae*	Protein Deletions/Swaps
3852	pTDH1-TDH1	pTDH2-TDH2	pTDH3-TDH3-Scarlet	*S. cerevisiae*	Scarlet Localization
3867	pTDH1-TDH1	pTDH2-TDH2	pTDH3-TDH1-Scarlet	*S. cerevisiae*	Scarlet Localization
3920	pTDH1-TDH1-Scarlet	pTDH2-TDH2	pTDH3-TDH3	*S. cerevisiae*	Scarlet Localization
3972	pTDH1-TDH1	pTDH2-TDH2-Scarlet	pTDH3-TDH3	*S. cerevisiae*	Scarlet Localization
3983	pTDH1-TDH1	pTDH2-TDH2-Scarlet	pTDH3-TDH3-Tq2	*S. cerevisiae*	Triple fluorophore
3984	pTDH1-TDH1-YFP	pTDH2-TDH2-Scarlet	pTDH3-TDH3-Tq2	*S. cerevisiae*	Triple fluorophore

Table 2: Primers used for gRNA hybridization

gRNA target	Sequence
TDH1	F: 5'-GATCTAAATAAAGCAATCTTGATGGTTTTAGAGCTAG-3' R: 5'-CTAGCTCTAAAACCATCAAGATTGCTTTATTTA-3'
TDH2	F: 5'-GATCTAAATCATTAAAGTAACTTAGTTTTAGAGCTAG-3' R: 5'-CTAGCTCTAAAACTAAGTTACTTTAATGATTTA-3'
TDH3	F: 5'-GATCACACACATAAACAAACAAAAGTTTTAGAGCTAG-3' R: 5'-CTAGCTCTAAAACTTTTGTTTGTTTATGTGTGT-3'
ΔTDH3	F: 5'-GATCAACAATGCAATAGCGCATCAGTTTTAGAGCTAG-3' R: 5'-CTAGCTCTAAAACTGATGCGCTATTGCATTGTT-3'

Table 3: Primers used for *TDH* deletion repair fragment generation

Primer Identity	Sequence
ΔTDH3-F	5'- CGGTAGGTATTGATTGTAATTCTGTAAATCTATTTCTTAAACTTCTTAAA TTCTACTTTTATAGTTAGTCCTTGAT-3'
ΔTDH3-R	5'- GTATCAGGTATCTACTACAGATATTACATGTGGCGAAAAAGACAAGAA CAATGCAATAGCGCATCAAGGACTAACTATAAAA-3'
ΔTDH2-F	5'- GTGTGTCTATTCATTTTATTATTGTTTGTTTAAATGTTAAAAAAACCAAG AACTTAGTTTCAAATTAAATTCATCACACAAACCATGCTCATGACATC-3'
ΔTDH2-R	5'- GGAATCTTTAATACATTTTCAATCTATTTAAGTTTTATAAACGTGTATAT GAGATGTCATGAGCATGGTTTGTGTGATG-3'
ΔTDH1-F	5'- GGTTTGATATTTCACCAACACACACAAAAAACAGTACTTCACTAAATTT

	ACACATAAAGCAATCTTGATGAGGATAATG-3'
ΔTDH1-R	5'-CTAGCAGAAAAACGGTAGTATTTATGTATATTCAAAAAAAAATCATTATCCTCATCAAGATTGCTTTATGTGTAAATTTAGTGAAGTACTG-3'

Table 4: Primers used for the amplification of *TDH* insertion repair fragments out of exogenous DNA sources

Primer Identity	Sequence	Exogenous DNA Source
TDH3-F	5'-GTTTTAAAACACCAAGAACTTAGTTTC-3'	*S. cerevisiae*
TDH3-R	5'-AAAATTTATTTAAATGCAAGATTTAAAGTAAATTCAC-3'	*S. cerevisiae*
TDH2-F	5'-CCAAGAACTTAGTTTCAAATTAAATTCATCACACAAACTTCGTACGCTGCAGGTCGAC-3'	*S. cerevisiae*
TDH2-R	5'-CGTGTATATGAGATGTCATGAGCATGCCGCGCGTTGGCCGATTCAT-3'	*S. cerevisiae*
TDH1-F	5'-CCAACACACACAAAAAACAGTACTTCACTAAATTTACACTTCGTACGCTGCAGGTCGACG-3'	*S. cerevisiae*
TDH1-R	5'-CCTCATCAAGATTGCTTTATTTAAGCCTTGGCAACATATTCGCCGCGCGTTGGCCGATTCAT-3'	*S. cerevisiae*
TDH1-YFP-F	5'-CGATGCCTCCGCTGGTATCC-3'	*S. cerevisiae*
TDH1-YFP-R	5'-GCCGACGTTCTCGCCATAAC-3'	*S. cerevisiae*
TDH3-Tq2-F	5'-AACGAATACGGTTACTCTACCAGAGTTGTCGACTTGGTTGAACACGTTGCCAAGGCTATGGTTAGTAAAGGTGAAGAATTGT-3'	pML-104 plasmid

TDH3-Tq2-R	5'-CTAAGTCATAAAGCTATAAAAAGAAAATTTATTTAAATGCAAGATTTAAAGTAAATTCACTTATTTATACAATTCATCCATACCT-3'	pML-104 plasmid

gRNA Construction Protocol: Link

CRISPR-Cas9 Yeast Transfection Protocol: Link