UNIVERSITY OF CALIFORNIA,
IRVINE

Engineering the oleaginous yeast *Yarrowia lipolytica* for industrial biosustainability

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Sciences

by

James Yu

Dissertation Committee:
Professor Suzanne Sandmeyer, Chair
Professor Haoping Liu
Professor Peter Kaiser
Professor Nancy Da Silva
Professor Aimee Edinger
Professor Qin Yang

2017

DEDICATION

In memoriam:

To CLY and SM, both of whom watched me start this journey but are not here to see me complete it.

"Do unto others 20% better than you would expect them to do unto you, to correct for subjective error."

-Linus Pauling

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would be remiss not to first thank my advisor and committee chair, Dr. Suzanne Sandmeyer. Her work ethic, positivity, and endless curiosity inspires everyone around her to be exemplary scientists and researchers. Indeed, the training I received these past seven years has not only helped me become a better scientist, but a better person overall. I believe that I am well prepared for whatever endeavor I plan on undertaking next, in no small part due to the training I've received. Also, I want to wish you all the best as you begin your new position as Senior Associate Dean of the School of Medicine. I have no doubt in my mind that you will excel in this new endeavor. The School of Medicine is very lucky to have you in your new capacity.

I would like to acknowledge both current and former Sandmeyer lab family members. Thank you to both Farbod Shavarebi and Dr. Virginia Bilanchone, not only for helpful discussions but also for putting up with my crazy antics and inane trash talking, especially in the later years of my doctoral work. Also, mad props for hanging on to the very end. To my former labmates Dr. Kristina Clemens, Dr. Kurt Patterson, and Dr. Ivan Chang: thank you for the support you've lended me these last several years. I hope to join you all on the other side, sooner rather than later. A special thank you goes out to my former lab mate Parth Sitlani. I'm so excited to be working with you once again.

I would like to thank all of my friends, both old and new, with whom I've spent countless hours venting on the phone, sharing memes on google hangouts, and typing inane (insane) emails back and forth. Without you all, I surely would have gone mad in grad school. Thank you to Arnold and Grant, who still want to see me every time I visit Rancho, even after suffering through two and a half decades of friendship with me. A

very special shoutout to my SD family (Jerry, Edith, Le, Jason, Jason, Mike, Josh, and respective SOs). And thank you to my newest forever friends that I've made here in Irvine: Selma, Stacey, Connie. Thank you for all the time you've spent at the pub with me, for being the harshest critics of both my science and my personal life, and for tolerating all the bitter beer I'm completely obsessed with. You all have helped me grow so much these last seven years.

I am also so thankful for my family. To my mom and dad, who came to America with nothing, yet were able to put three kids through college and still have enough in the bank to retire comfortably: I can't thank you enough for your support throughout the years. I hope everyone who reads this is as inspired as I am by your bravery, work ethic, and wisdom. To my younger siblings, Doris and Louis, both of whom are vastly more successful than their older brother: I aspire to be like you when I finally grow up.

Thank you to my committee members, Dr. Aimee Edinger, Dr. Haoping Liu, Dr. Peter Kaiser, Dr. Qing Yang, and Dr. Nancy Da Silva for all of your advice and helpful comments throughout this process. Even though you've repeatedly mentioned, "Oh but take my advice with a grain of salt, I'm not a yeast person", the insight into my project that you've all provided me truly invalidates that qualifier.

Lastly, thank you to the Center for Biorenewable Chemicals and all of the scientists and support staff at Iowa State University. Chapter 5 of this dissertation was a direct collaboration with Adam Okerlund's and George Kraus's groups and this work would have never been completed without their assistance and input. I also would like to acknowledge that this research was funded by CBiRC and the NSF (Grant no. EEC-0813570, Brent Shanks, PI; subcontract Suzanne Sandmeyer).

CURRICULUM VITAE

James Yu

EDUCATION

2017 **Ph.D. Biomedical Sciences**
Department of Biological Chemistry, UC Irvine
2007 **Teaching Credential**
University of California, San Diego
2007 **M. Ed, Secondary Education**
University of California, San Diego
2006 **B.S. Cell Biology/Biochemistry**
Department of Biology – University of California, San Diego

RESEARCH AND PROFESSIONAL EXPERIENCE

2010-2017 **Graduate Student** – Department of Biological Chemistry, UC Irvine, CA
Advisor: Suzanne Sandmeyer
Thesis: Engineering the oleaginous yeast, *Yarrowia lipolytica*, for industrial biosustainability
2006-2010 **Science Teacher –** San Diego Unified School District, San Diego, CA
2004-2006 **Undergraduate Researcher** - Department of Psychology, UCSD, CA

PUBLICATIONS

Magnan C*, **Yu J***, Chang I*, et al. (2016) Sequence Assembly of *Yarrowia lipolytica* Strain W29/CLIB89 Shows Transposable Element Diversity. PLOS ONE 11(9): e0162363.

Patterson K*, **Yu J***, Landberg J, et al. Functional genomics for the oleaginous yeast *Yarrowia lipolytica*. (in revision)

Yu J, Landberg J, et al. Production of triacetic acid lactone in the oleaginous yeast *Yarrowia lipolytica*. (in preparation)

*indicates co-first authorship

AWARDS

2015 **Yu J., Zhao L., "Optimizing TAL production in *Yarrowia lipolytica* - A minimalist approach" (2015)**
Student Led Research Grant (CBiRC)
Amount awarded: $9,500

2014 **Chang I., Yu J., Chowdhury A., "Predicting Genomic Interventions for Lipid Overproduction in *Yarrowia lipolytica*"**
Student Led Research Grant (CBiRC)
Amount awarded: $5,000

2014 **Designer Metabolism: Polyketide production in an oleaginous yeast**
CBiRC Perfect Pitch Contest
2nd place

2013 **Chang I., Yu J., Tee T., "13C metabolic flux analysis based fluxomic comparison between *Saccharomyces cerevisiae* and oleaginous yeast *Yarrowia lipolytica"* (2013)**
Student Led Research Grant (CBiRC)
Amount awarded: $10,000

2012 **Thomas P., Yu J., Suhalim J., "The effects of nitrogen limitation on the distribution of lipid accumulation in yeast" (2012)**
Center for Complex Biological Systems (UCI) Opportunity Award
Amount awarded: $10,000

2012 **Lim S., Chang I., Yu J., "Studying the role of respiration and mitochondrial dynamics in the morphology of *Yarrowia lipolytica"* (2012)**
Center for Complex Biological Systems (UCI) Opportunity Award
Amount awarded: $10,000

SERVICE AND LEADERSHIP

2015 **Student Leadership Council Student Advisor**
Center for Biorenewable Chemicals (CBiRC)
Iowa State University

2014 **Student Leadership Council Chair**
Center for Biorenewable Chemicals (CBiRC)
Iowa State University

2012 **Student Leadership Council Secretary**
Center for Biorenewable Chemicals (CBiRC)
Iowa State University

2011 **Graduate Student Representative**
Department of Biological Chemistry
University of California, Irvine

2008 **Junior Class Advisor (Class of 2010)**
San Diego High School
San Diego, CA

INVITED TALKS

May 2015 CBiRC 7th Annual Site Visit

"2015 Student SWOT Results"
Iowa State University

Oct 2014 **NSF ERC Meeting – Perfect Pitch Competition**
"Yarrowia lipolytica – A next generation microbial biofactory"
Washington DC

May 2014 **CBiRC 6th Annual Site Visit**
"2014 Student SWOT Results"
Iowa State University

ABSTRACT OF THE DISSERTATION

Engineering the oleaginous yeast, *Yarrowia lipolytica*, for industrial biosustainability

By

James Yu

Doctor of Philosophy in Biomedical Sciences

University of California, Irvine, 2017

Professor Suzanne Sandmeyer, Chair

The conversion of renewable biomass to industrial chemicals of interest through microbial synthesis represents a clean sustainable alternative to the finite supply of dirty fossil fuels that modern society is heavily dependent on. In particular, one class of biological molecules with industrial potential are lipids. Lipids and their precursors are suitable for conversion into a diverse array of industrial chemicals such as polymers, biofuels, and antibiotics. Furthermore, a subset of microbes, known as oleaginous yeast, are capable of accumulating high levels of intracellular lipids. However, the paucity of background knowledge and molecular tools of oleaginous yeast impedes the widespread adoption of these organisms by the biorenewable chemical industry. To address this shortcoming, we apply both familiar molecular cloning techniques along with innovative genomic technologies to greatly expand the molecular and genomic toolkit of the oleaginous yeast *Yarrowia lipolytica*. We publish the most complete genome assembly of *Y. lipolytica* and utilized a transposon mutagenesis approach to query the genome and assign essentiality classifications to every annotated gene. We then utilize these tools to engineer a novel microbial platform for the synthesis of pogostone, a valuable antibiotic, from renewable glucose. This work not only features the development of a pipeline detailing the biorenewable production of a specific value-added chemical, but it also broadly advances the field of study pertaining to the oleaginous yeast *Y. lipolytica*.

CHAPTER 1: Introduction

The Industrial Revolution marked the beginning of the “Age of Oil” in modern society. Petroleum provides power to our homes and workplaces and fuels our cars and airplanes. Modern polymers and plastics synthesized from fossil fuels are used to make everything from the cars we drive to the clothes we wear. In addition to the negative ecological impacts associated with the burning and mining of petroleum, one of the biggest problems with modern society’s reliance on fossil fuels is its inherent limited availability (Shafiee and Topal 2009; Hook and Tang 2013). Fossil fuels are a nonrenewable resource; once depleted, fossil fuels cannot be regenerated in a practical time scale. Therefore, a sustainable alternative to petroleum and petroleum derived chemicals must be achieved. One potential solution is to use microbes such as yeast, bacteria, and algae as microbial hosts to produce chemicals that can either supplement or replace existing fossil fuel technology from renewable carbon sources. The ubiquitous model laboratory bacterium *Escherichia coli* has been leveraged to produce 1,3 propanediol and 1,4 butanediol, two hydrocarbon compounds that can be used to synthesize plastics and polymers from renewable glucose (Nakamura and Whited 2003; Yim, Haselbeck et al. 2011). In 2000, DuPont successfully commercialized this technology into the Sorona line of polyester fiber (Kurian 2005). Similarly, the most widely studied yeast species in biology, *Saccharomyces cerevisiae*, is responsible for the majority of the world’s production of ethanol, which is used in solvents, alcoholic beverages, and as a ubiquitous supplement to commercial automobile gasoline (Lopes, Paulillo et al. 2016). Similar to DuPont’s synthesis of diols from *E. coli*, the ethanol produced by *S. cerevisiae* is also sourced from glucose and other renewable sugars. Both of these two examples demonstrate that microbial-derived, renewably sourced petrochemicals are both technologically and commercially feasible in this present day and age.

The main advantage of engineering both *E. coli* and *S. cerevisiae* for the production of renewable chemicals is the abundance of tools, techniques, complete genome assemblies, and accumulated knowledge as a result of the decades scientists have spent studying these two microbes. Although well suited for the production of the chemicals mentioned above, these two model microbes are not ideal hosts for the production of all the chemicals and fuels derived from petroleum. For instance, lipids and fatty acids serve as precursors to biodiesel and a broad range of different detergents, lubricants, and polymers (Zhu and Jackson 2015). However, lipid titers in unmodified *S. cerevisiae* strains typically range from 4-5% dry cell weight (Runguphan and Keasling 2014). Even the most heavily modified *S. cerevisiae* strains only reach 45% lipid (Kamisaka, Kimura et al. 2013; Zhou, Buijs et al. 2014).

S. cerevisiae and other Crabtree positive yeast have evolved a unique metabolism that initially ferments glucose into ethanol, even in the presence of oxygen (Christen and Sauer 2011; Pfeiffer and Morley 2014). Efforts to utilize this organism as a host microbe for the production of ethanol leverage this inherent propensity to drive glucose flux toward ethanol. Furthermore, *S. cerevisiae* also has a tendency to store excess carbon in the form of complex carbohydrates like trehalose and glycogen (Nissen, Schulze et al. 1997; Christen and Sauer 2011). The tendencies of *S. cerevisiae* to efficiently produce ethanol and also store starch and other carbohydrates rather than lipids overshadows the convenience of this well characterized model yeast as a host for the production of lipids and related compounds.

In contrast, in conditions of nitrogen starvation, some oleaginous yeast species are capable of accumulating up to 70% dry cell weight in lipids without any additional genetic modifications (Li, Zhao et al. 2007; Dong, Knoshaug et al. 2016). These oleaginous yeast, which are inherently capable of accumulating high levels of lipids, have recently emerged as potential alternative host organisms for the biorenewable production of select chemicals. In particular, the long hydrophobic hydrocarbon chains that result from fatty acid synthesis closely resemble the petroleum based hydrocarbons that comprise the prototypical fossil fuel. Indeed, microbial oils can be converted into biodiesel, and represent a promising renewable fuel source (Li, Du et al. 2008).

Out of the many species of oleaginous yeast, *Yarrowia lipolytica* is among the best characterized (Beopoulos, Cescut et al. 2009). Although lipid accumulation in *Y. lipolytica* is not very high compared to other oleaginous organisms (18% dry cell weight in lipids), *Y. lipolytica* has emerged as one of the most commonly used oleaginous yeasts, due to the early availability of a molecular “toolkit” (Wolf 1996). Despite not being as complete and ubiquitous as the molecular “toolkits” available for the model laboratory microbes *E. coli* and *S. cerevisiae*, the limited plasmids and protocols that have been developed to enable the manipulation of metabolic and genetic networks has brought *Y. lipolytica* to the forefront as the model oleaginous yeast for diverse biotechnological applications. In addition to being leveraged for the production lipids and fatty acids, *Y. lipolytica* has also been utilized as a host microbe for the production of an array of chemicals including fatty acids, organic acids, and polyols (Tomaszewska, Rywinska et al. 2012; Liu, Chi et al. 2013; Yuzbasheva, Mostova et al. 2017; Zeng, Zhang et al. 2017). In contrast, there are only a handful of studies that have investigated the metabolic engineering of the oleaginous yeasts *Rhodotorula glutinis* and *Lipomyces starkeyi*, which may be attributed to the fact that protocols for genetic manipulation for these two yeasts have only been developed within the last four years (Li, Sun et al. 2013; Calvey, Willis et al. 2014).

All of these oleaginous yeast species demonstrate high levels of *de novo* fatty acid and lipid synthesis. In fatty acid synthesis, cytosolic malonyl CoA molecules are iteratively condensed onto a cytosolic acetyl CoA starter molecule, adding two carbons at a time to the growing hydrocarbon chain. Once sufficient length is achieved, the growing chain is terminated by a thioesterase and released from the enzyme. Ultimately, since malonyl CoA is derived from the carboxylation of cytosolic acetyl CoA through the action of acetyl CoA carboxylase, high levels of fatty acid synthesis requires high carbon flux into pathways to produce cytosolic acetyl CoA. One unique distinction of oleaginous yeasts that can accumulate high levels of *de novo* lipids compared to fermentative yeasts is the existence of the ATP Citrate Lyase (ACL) pathway. This pathway has been suggested to be the primary source of cytosolic acetyl CoA in conditions of nitrogen limitation and carbon excess (Boulton 1981; Blazeck, Hill et al. 2014; Pomraning, Bredeweg et al. 2017). In conditions of nitrogen limitation, oleaginous yeasts begin to metabolize AMP as a source of nitrogen. Since AMP is an allosteric activator of isocitrate dehydrogenase, this decrease in intracellular AMP levels inhibits the activity of the enzyme, stopping further progress into the TCA cycle, and leading to an accumulation of isocitrate and citrate in the mitochondria. The accumulated citrate is transported into the cytosol, where it serves as a substrate for ACL to produce cytosolic acetyl CoA. It is believed that the redirection of carbon flux from mitochondrial citrate and the TCA cycle into cytosolic acetyl CoA and fatty acid synthesis is responsible for the increased lipid accumulation in nitrogen limiting conditions.

Despite the work contributed by many industry and academic groups to the study of *Y. lipolytica*, the knowledge of its metabolism and genetics still pales in comparison to biologists' understanding of *S. cerevisiae* and *E. coli*. In order to address the relative paucity of background information in this yeast, the first part of this dissertation (chapters 2-4) describes the work we have done to contribute basic genetic and genomic tools to advance the *Y. lipolytica* engineering field by expanding the molecular "toolkit". The latter part of this dissertation (chapter 5) describes the first study that utilizes *Y. lipolytica* as a host organism for production of the polyketide triacetic acid lactone (TAL). Like lipids, TAL is also an acetyl CoA based metabolite and several groups have shown that it can be differentiated into a broad range of commodity and value-added chemicals (Chia, Haider et al. 2013; Kraus, Wanninayake et al. 2016).

Chapter 2 describes the molecular tools and techniques that we developed that laid the foundation for our molecular work in this organism. We utilized homologous recombination to construct *ura3* and *leu2* mutant strains and were able to clone the respective genes onto plasmids to utilize as auxotrophic selection markers. These strains and plasmids were eventually utilized to characterize promoter activity using a green fluorescent protein based assay coupled with flow cytometry. Finally, we also optimized a transformation protocol to maximize efficiency of transformation into *Y. lipolytica*. None of the later molecular work would have been possible without the expression plasmid system developed in this chapter. The Hermes mutagenesis work described in chapter 4 was heavily reliant on the optimized transformation protocols detailed in this chapter.

Chapter 3 details our work to assemble the most complete genome assembly and annotation of *Y. lipolytica* to date. Although a reference genome has been assembled and is publicly available, it was built in 2004, using the best technology at the time. Consequently, this reference genome assembly is incomplete and composed of multiple contigs. Furthermore, this genome assembly is of CLIB150, a strain that is not presently used for industrial applications. In order to address these shortcomings, we undertook a hybrid sequencing approach that combined both long PacBio sequencing reads with high quality Illumina sequencing reads to assemble the genome of CLIB89, an industrially relevant strain of *Y. lipolytica*. This strategy successfully assembled the genome into chromosome-length contigs with minimal errors. Even repetitive elements, which are usually difficult to place accurately in the genome, were successfully assembled in our build and, as a result, we were able to characterize a new Ty3/gypsy like retroelement in *Y. lipolytica*. Afterwards, several open-source software packages were used to predict open reading frames and annotate the genes in this assembly.

Chapter 4 describes work we did to functionally characterize this new genome assembly. The Hermes DNA cut-and-paste transposon was used to perform saturation mutagenesis of the of *Y. lipolytica* genome. The Hermes element A Hermes mobilization into a gene would disrupt its reading frame, thus we assumed that genes with a low frequency of Hermes insertions would be essential for growth. We use a Poisson based probabilistic model to classify every gene in the *Y. lipolytica genome* as essential, nonessential, or conditional for growth in glucose and glycerol by the number of Hermes insertions in and around each gene. Several positive outcomes with significant benefit to the *Y. lipolytica* community resulted from this work. The list of essential genes resulting from this study would inform and streamline future metabolic engineering efforts through both rational pathway engineering and also through more precisely refined *in silico* genome scale metabolic models. Additionally, this study also generated a diverse collection of mutants which can be used to screen for favorable industrial phenotypes. We use BODIPY, a lipophilic dye coupled with fluorescence activated cell sorting to isolate mutants that accumulate great levels of lipids.

Chapter 5 describes our efforts in engineering *Y. lipolytica* for the production of triacetic acid lactone (TAL), an industrially relevant polyketide metabolite. As mentioned above, *Y. lipolytica* is already being utilized for the production of diverse industrially relevant compounds such as organic acids and polyols, in addition to the lipids commonly associated with oleaginous yeasts. Although these secondary metabolites do not resemble the long hydrocarbon chains that are characteristic of fatty acids and lipids, all of these metabolites are derived from the same acetyl CoA that is necessary for fatty acid synthesis. Similarly, TAL is a polyketide synthesized from acetyl CoA and malonyl CoA, and thus we describe how this oleaginous yeast may serve as a good host for the production of these molecules. This chapter describes work to engineer fermentation conditions for the production of TAL in a minimal medium formulation. We show that this minimal medium formulation is not only suitable for cultivating robust growth, we also show that it is amenable for economical production and purification of TAL for the purpose of downstream derivation into pogostone, a high value chemical.

CHAPTER 2: Optimization of non-integrative and integrative transformations in *Yarrowia lipolytica*

Introduction

The yeast *Yarrowia lipolytica* possesses a number of characteristics that differentiate it from model yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. *Y. lipolytica* efficiently metabolizes potentially low-value or waste hydrophobic substrates, fatty acids and complex hydrocarbons like paraffin. On the production side, it is not only competent for large-scale secretion of proteins but as an oleaginous yeast (Ogrydziak and Scharf 1982; Barth and Gaillardin 1997; Madzak, Nicaud et al. 2005; Beopoulos, Chardot et al. 2009; Gasmi, Ayed et al. 2011), it can generate a high percentage of mass in lipid or in intermediates such as citrate and fatty acids for industrial applications including renewable dietary additives, detergents, synthetic fuels, and environmental remediation (Bankar, Kumar et al. 2009; Coelho 2010) . Compared to other so-called "non-conventional" yeasts, *Y. lipolytica* also has advantages including an annotated genome, marker genes, and selectable plasmids. However, one current limitation of *Y. lipolytica* with respect to biotechnological applications is the dominance of non-homologous end joining (NHEJ) repair over homologous recombination (HR), resulting in low frequencies of targeted integration, in turn making genomic modification cumbersome (Davidow, Apostolakos et al. 1985; Xuan, Fournier et al. 1988; Fickers, Le Dall et al. 2003; Wang, Hung et al. 2011).

We first sought to increase the efficiency of transformants recovered overall by optimizing transformation beginning with several previous approaches. We applied this protocol to demonstrate that *Y. lipolytica* exhibits a population, which is transformation competent so that co-transformation can be used to further leverage available markers. Comparison of the strains lacking the NHEJ factors Ku70 and Pol4, using the optimized transformation method showed an elevated proportion of targeted integration in the *ku70*Δ strain accompanied by a decrease in overall integration, but this was not observed in the Pol4 deficient strain, suggesting that if this gene is involved in NHEJ in *Y. lipolytica*, it is redundant with other factors.

Results

To evaluate the transformation capabilities of *Y. lipolytica*, the effectiveness of three established heat-shock based protocols for CEN/ARS plasmid transformations was compared. Protocol #1 is a *S. cerevisiae* transformation method that has been adapted for use with *Y. lipolytica* (Barth 1996; Kretzschmar, Otto et al. 2013). Protocol #2 is an optimized version of the *Y. lipolytica* protocol established by Davidow *et al (Chen, Beckerich et al. 1997).* Protocol #3 is a high-efficiency method developed for *S. cerevisiae* (Gietz and Schiestl 2007) (Figure 2.1). In each case, competent *ura3*Δ cells (yJY1832) were transformed with a *URA3*-marked plasmid (pJY3731) and the transformation efficiencies of each protocol were determined (Figure 2.2A). Each protocol produced between 1 to 3 x 10^5 transformants per μg plasmid DNA (0.5 to 1.5 x 10^6 transformants per ρmol plasmid DNA) which is similar to or higher than those published in the literature for *Y. lipolytica*, which range from 10^3 to 10^5 transformants per μg plasmid DNA (Barth 1996; Chen, Beckerich et al. 1997). The results were less than the highest high-efficiency *S. cerevisae* protocol, which were as high as 10^6

transformants per µg plasmid DNA (Gietz and Schiestl 2007). Transformations with protocol #3 consistently produced two-fold higher efficiencies relative to the other two protocols tested, and so this protocol was the focus of further work.

Protocol #3 was developed as a quick means to rapidly perform a large number of transformations and involves the addition of a single buffer to the competent cells, immediately followed by heat shock. In order to optimize this protocol, a number of specific modifications were tested. Using protocol #3 as the basis for each transformation, the effect of PEG of different molecular weight (PEG-3350 versus 4000), heat shock temperatures (39°C versus 42°C), heat shock duration (40 versus 60 m), OD_{600} of competent cells at the time of harvest (OD_{600} of 1.5 versus 5.0 OD), time of incubation in LiOAc prior to transformation, and an overnight 4°C incubation of the transformed cells versus plating immediately following heat shock, were evaluated. Based on the efficiencies generated from each variation of protocol #3, it was clear that the majority of the alterations had little effect on *Y. lipolytica* transformation (Figure 2.2B). However, one modification that did significantly increase the transformation efficiency was the overnight 4°C incubation of heat shocked cells prior to plating, which caused a three-fold increase in the number of transformed cells. Although the possibility cannot be formally excluded that transformed cells replicated as a subpopulation, there was not a gross change in the number of cells during the overnight incubation, as the OD_{600} remained constant between heat shock and plating (data not shown). The resultant transformation efficiency is higher than previously published *Y. lipolytica* protocols, and therefore this modification of protocol #3 was utilized for the remainder of the study.

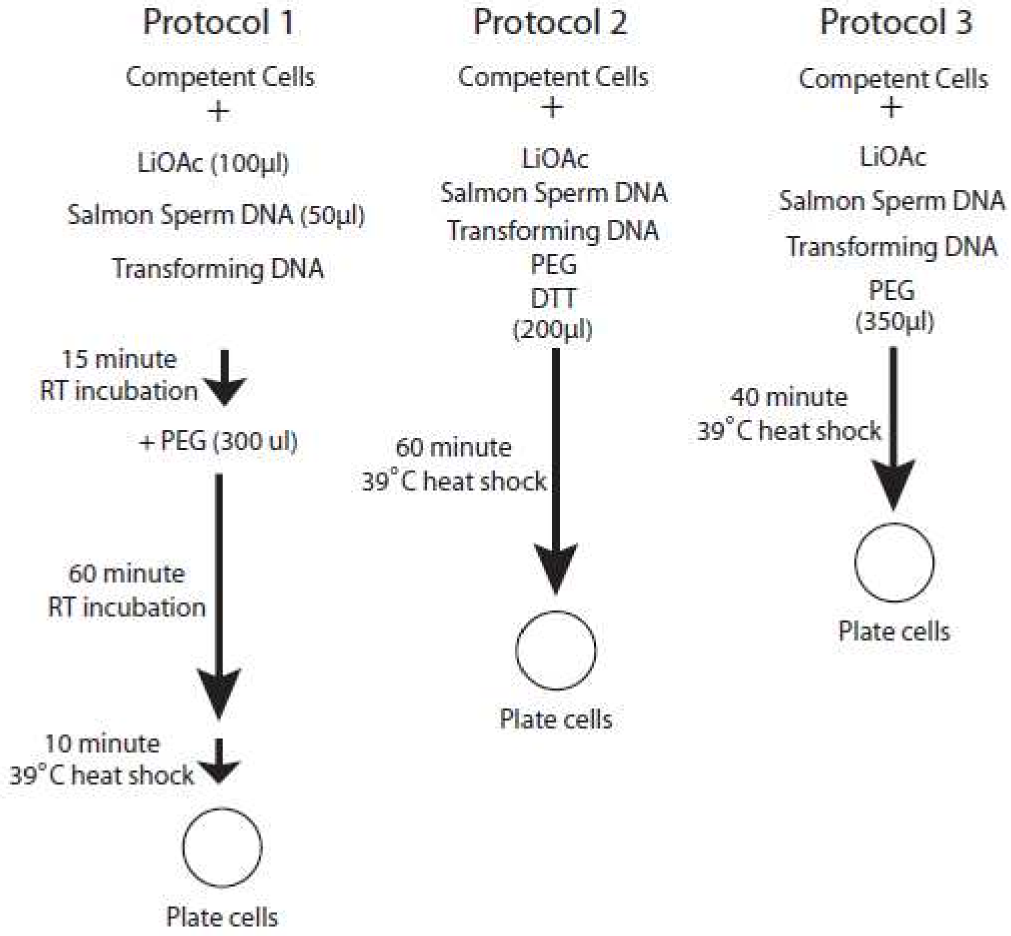


Figure 2.1. Development strategy for transformation protocols. The reagents added and specific incubation/heat shock times are indicated.

In *S. cerevisiae*, a minority of cells within the population have enhanced competency to take up foreign DNA, which is useful in situations where it is difficult to mark the DNA of interest (Gietz and Schiestl 1991). We examined whether *Y. lipolytica* shares this property by measuring the frequency of co-transformation. A *ura3*Δ *leu2*Δ strain was transformed under the optimized transformation conditions, individually with a *URA3*-marked plasmid, a *LEU2*-marked plasmid or both simultaneously and plated on the appropriate selection media. The transformation efficiencies were similar for each plasmid, and only a five-fold decrease in co-transformation efficiency of both plasmids was observed (Figure 2.3). This is similar to what has been documented in *S. cerevisiae*, which exhibits a three to ten-fold drop in co-transformation efficiency compared to single transformation efficiency (Hicks, Hinnen et al. 1979; Gietz and Schiestl 1991). Roughly 0.1% of cells are transformed per □g of single plasmid (transformation efficiency of 1 x 10^5 when utilizing 1 x 10^8 competent cells) (Figure 2.2A). If plasmid uptake were completely random, only 0.0001% of cells would acquire both plasmids, rather than the 0.02% observed in the co-transformation experiment, a significant increase. This indicated that there is a fraction of cells which have enhanced competency for DNA uptake.

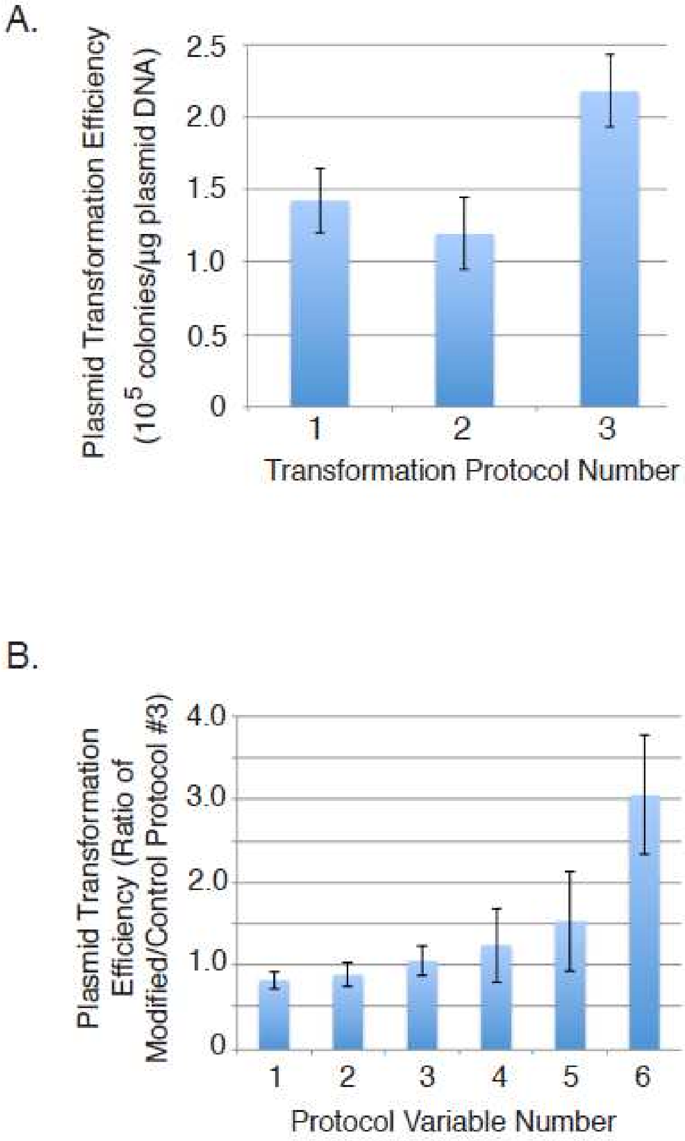


Figure 2.2. Comparison of transformation protocols and variables.
(A) Transformation of the *ura3*Δ strain (yJY1832) with a *URA3*-marked plasmid (pJY3731) using three different protocols was performed and the transformation efficiencies (number of colonies/µg plasmid DNA) were calculated. Each transformation was performed at least three times and the standard error of the mean (SEM) is shown. **(B)** Protocol #3 was performed to individually test six variables – (1) transformation mix with PEG4000; (2) 42°C heat shock; (3) 60 m heat shock; (4) competent cells harvested at 5 OD; (5) competent cells incubated in LiOAc buffer overnight before transformation; and (6) heat-shocked cells incubated at 4°C overnight before plating. The transformation efficiencies of the modified protocols relative to the control protocol are indicated.

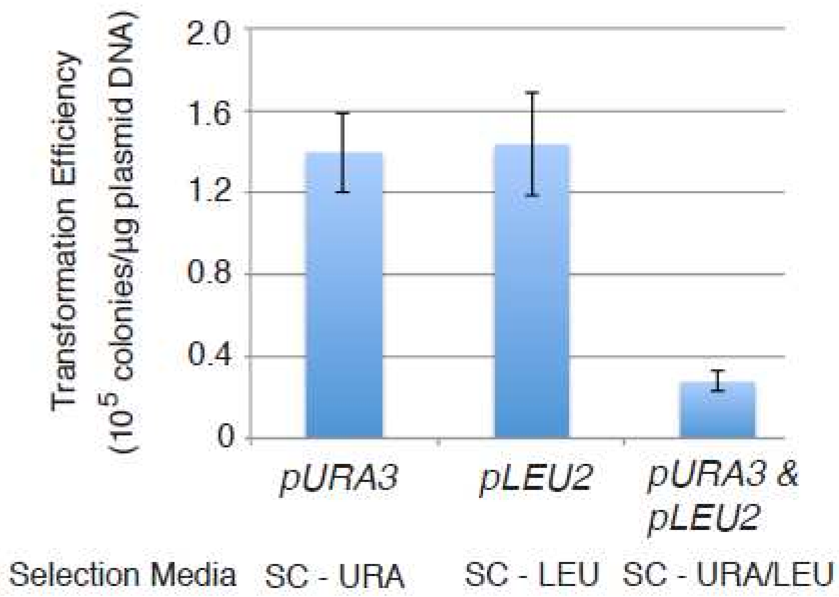


Figure 2.3. Co-transformation frequency in *Y. lipolytica*. The *ura3*Δ *leu2*Δ strain (yJY1897) was transformed individually with a *URA3*-marked *ars18* plasmid (pJY3731), a *LEU2*-marked *ars18* plasmid (pJY3830) or both plasmids. Transformations were plated on the appropriate selection media and transformation efficiencies (number of colonies/µg plasmid DNA) were calculated.

Optimal transformation efficiency is obviously an important consideration for engineering strains, but for *Y. lipolytica* bioengineering a secondary limitation is often the difficulty of targeting integrations to a specific genomic site. Therefore, the parameters affecting targeted integrative transformations were evaluated. In *S. cerevisiae*, which famously has very efficient HR repair, genomic integration can be targeted with as little as 50bp of terminal homology (Baudin, Ozier-Kalogeropoulos et al. 1993; Oldenburg, Vo et al. 1997). However, *Y. lipolytica* generally requires more than 750bp of terminal homology for significant efficiency (Fickers, Le Dall et al. 2003; Kretzschmar, Otto et al. 2013).

To determine the relationship between length of homology and efficiency of targeted integration for the optimized transformation protocol, the fraction of targeted integration versus untargeted integration at two common marker loci, *URA3* and *LEU2,* was determined. *Ura3Δ* (yJY1832) was transformed with *URA3* cassettes having ~200, 500 and 1000 bp flanking the region deleted in yJY1832 and the frequency of transformation was determined. Ura$^+$ cells would arise regardless of whether the cassette was integrated via HR or NHEJ. Ura$^+$ cells were produced with an integrative transformation efficiency of 10^4 colonies per ρmol DNA (Figure 2.4A). Length of flanking sequence did not have a consistent effect on transformation efficiency (Figure 2.4A). We next examined the frequency of *URA3* knockout in wt CLIB89 by transforming with an analogous *URA3* knockout cassette with varying lengths of flanking sequence. These were the same as those used to integrate *URA3* into the genome except lacking the *URA3* marker sequence. Transformants were plated on YPD and transferred to SC containing 5' fluoro-orotic acid (5FOA). Because 5FOA is toxic to *URA3* cells, only transformants in which the *URA3* knockout cassette replaced the locus by HR would survive on 5FOA. The frequency of *URA3* knockout and thus targeted integration increased with flanking sequence (Figure 2.4B). The effect of the length of flanking sequence homology on targeting was determined by dividing the frequency of 5FOA-resistant transformants by total Ura$^+$ transformants in the *ura3*Δ background with 200, 500, and 1000bp of flanking sequence. The proportion of targeted integration increased from 3% with 200bp homologous sequence to over 10% with 1000 bp of flanking sequence (Figure 2.4C). Because other studies have shown locus-specific differences (Fickers, Le Dall et al. 2003), a *URA3* cassette flanked by sequence targeting the cassette to *LEU2* was also tested for the effect of homologous flanking sequence length on targeted transformation. This *URA3* cassette contained the region that was deleted in the *ura3Δ* (yJY1832) strain, thus removing the possibility of targeting the cassette to the *URA3* locus via HR. As observed at the *URA3* locus, more extensive homology increased the proportion of HR events with Ura$^+$ Leu$^-$ transformants representing less than 5% of total Ura$^+$ transformants when transformed with a cassette containing 200bp *LEU2*-flanking sequences compared to 20% with 1000bp *LEU2* flanking sequences (Figure 2.4D). Thus, HR, which is not the dominant means of integration in in *Y. lipolytica,* is clearly enhanced with longer homologous flanking sequence.

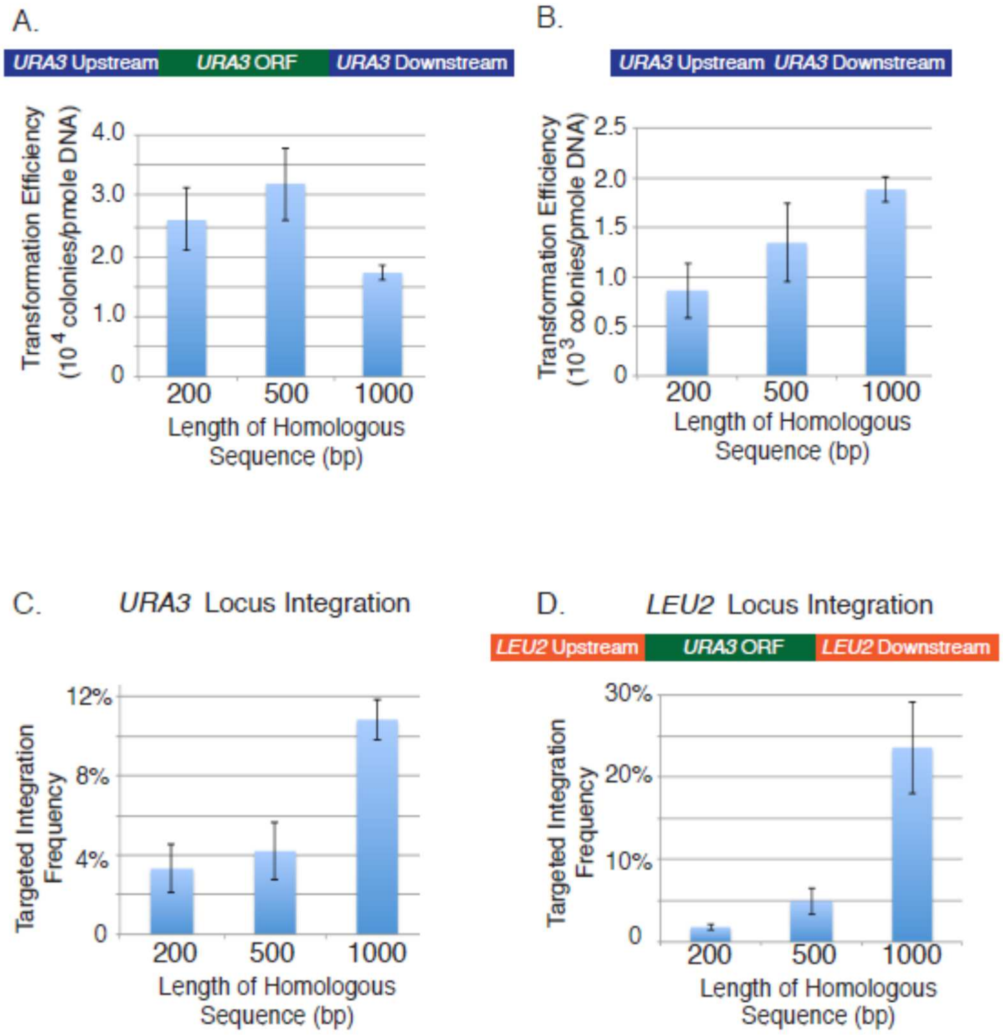


Figure 2.4. Determination of the frequency of sequence-specific gene integration. The transformation efficiencies (number of colonies/ρmole DNA) are shown for *URA3* knockin and *LEU2* knockout transformations. **(A)** The *ura3*Δ *st*rain was transformed with a *URA3* cassette containing *URA3* 5' and 3' flanking sequence of the indicated length and the *URA3* ORF. Cells were plated on uracil-minus media. **(B)** A wt strain was transformed with a *URA3* knockout cassette containing *URA3* 5' and 3' flanking sequences of the indicated length. Cells were plated on YPD and then replica plated on 5-FOA. **(C)** The ratio of *URA3* knockout versus knockin is shown to illustrate the ratio of specific integration to total integration. **(D)** A *ura3*Δ strain was transformed with a *URA3* cassette containing *LEU2* 5' and 3' flanking sequences of the indicated length. Cells were plated on SC minus uracil and then replica plated onto SC minus leucine to identify specific integrants. The percentage of transformed cells that were *leu2*Δ is indicated. Over 500 Ura$^+$ colonies were replica plated with the above integration cassettes to determine the targeted integration frequency.

NHEJ factors have been described in *S. cerevisiae* for their role in repair of double-stranded DNA breaks. In *S. cerevisiae*, components of NHEJ include Ku70 *and* Ku80, which recognize free DNA ends; Pol4, the Pol IV DNA repair polymerase; and Dnl4, Lif1 and Nej1, members of the DNA ligase complex. The *Y. lipolytica* genome was examined for both protein and gene homologs to these factors using the BLASTP and BLASTN algorithms respectively (Sherman, Martin et al. 2009). Ku70 and Ku80 homologs identified were consistent with those previously reported [13, 25]. *Ku70*Δ and *ku80*Δ strains have been previously examined for their ability to specifically target foreign DNA with one report showing little effect of *ku80*□ [13,26]. In addition, a locus was identified with similarity to *Mus musculus* Pol λ, a mammalian Pol IV protein (Sherman, Martin et al. 2009). BLAST analysis failed to identify expected orthologs of *LIF1* and *NEJ1*. Efforts to delete *DNL4* were unsuccessful. Therefore, *KU70,* and *POL4* were individually deleted as described in the Materials and Methods yielding strains yYJ1863 and yJY1900, respectively. These strains were examined for the frequency of Ura+ Leu$^-$ cells resulting from transformation with the *URA3* cassette targeting the *LEU2* locus. *Ku70*Δ cells showed targeted integration of the cassette in roughly 50% of transformants compared to wt cells at about 20% (Figure 2.5A). In contrast to the *ku70*Δ strain, the *pol4*□ strain did not show enhanced frequency of targeted integration (Figure 2.5A). Consistent with these results, the overall integrative transformation frequency decreased in *ku70*Δ cells which were defective for the major pathway, but was unchanged in *pol4*Δ strains which did not exhibit an increase in HR (Figure 2.5B). Overall plasmid transformation efficiency was not affected (Figure 2.5C).

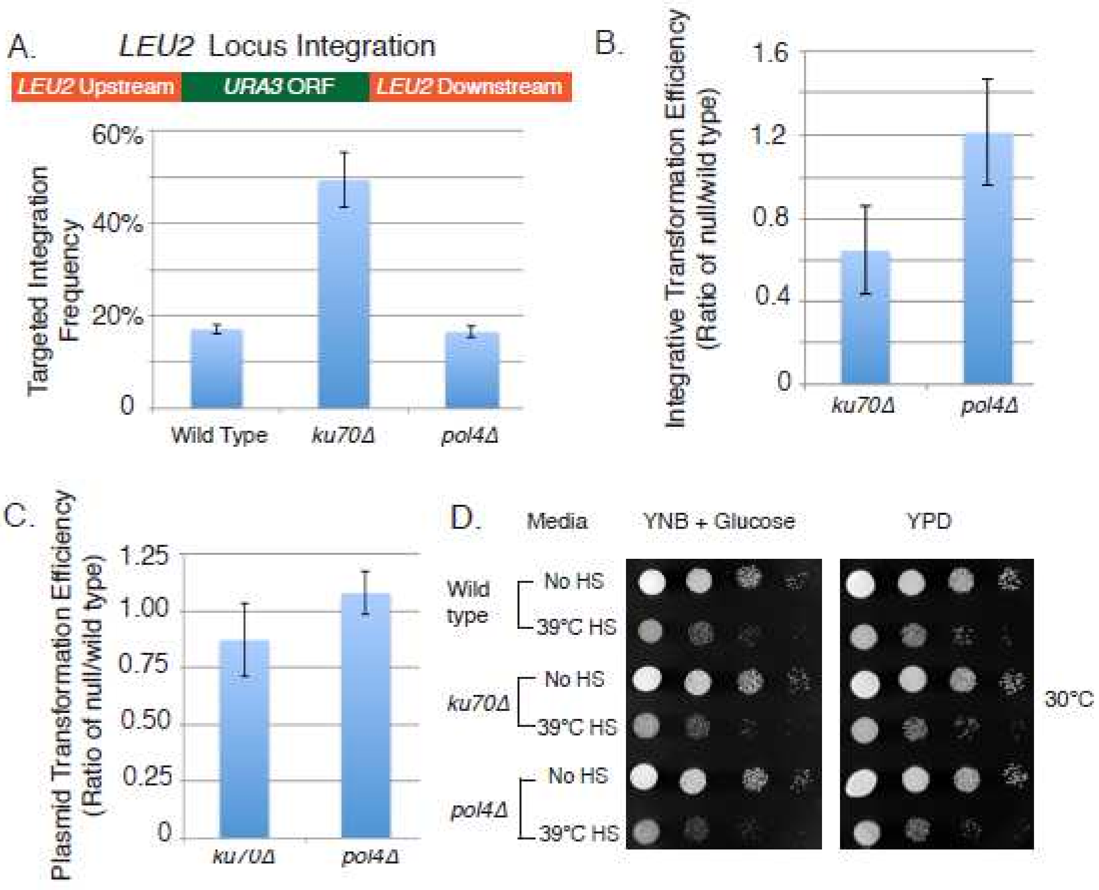


Figure 2.5. Examination of targeted integration and cell viability in NHEJ mutants. **(A)** *Ura3*Δ, *ku70*Δ *ura3*Δ (yJY1863), and *pol4*Δ *ura3*Δ (yJY1900) strains were transformed with a *URA3* cassette containing 1000bp *LEU2* 5' and 3' upstream and downstream sequences surrounding the *URA3* ORF. Cells were plated on SC minus uracil and then replica plated onto SC minus leucine to confirm specific integration. The percentage of transformed cells that were *leu2*Δ is indicated. Multiple *ku70*Δ *ura3*Δ and *pol4*Δ *ura3*Δ isolates and over 500 Ura$^+$ colonies were replica plated for each null to determine the targeted integration frequency. **(B)** The transformation efficiencies for the experiment in **(A)** were calculated. The transformation efficiencies for the *ku70*Δ and *pol4*Δ strains were normalized to the wt strain. **(C)** Plasmid transformation efficiencies for wt, *ku70*Δ and *pol4*Δ strains were calculated. Mutant transformation efficiencies were normalized to the wt strain. (D) Wt, *ku70*Δ (yJY1862) and *pol4*Δ (yJY1886) strains were harvested in log phase and spotted onto YPD or minimal medium. Each spot is a ten-fold serial dilution of the previous spot. A fraction of the cells were subjected to a 40 m 39°C heat shock. Plates were incubated for~24 h at 30°C.

Because the Ku complex contributes to stress resistance in *S. cerevisiae*, and industrial conditions can involve growth at elevated temperatures, we examined *ku70*Δ and *pol4*□ strains for growth on YPD and YNB dextrose media after being subjected to heat stress by a dilution spotting assay (Figure 2.5D). Although growth was decreased after elevated temperature stresses, no significant difference was observed between the wt and *ku70Δ* and *pol4Δ* mutants.

Discussion

Y. lipolytica has multiple features contributing to its importance as an industrial microbe. In this work, we sought to increase the efficiency of genetic modifications required for its bioengineering. A transformation protocol developed for *S. cerevisiae* [20] was optimized for *Y. lipolytica* and shown to significantly improve transformation efficiency relative to other published protocols and achieve up to 10^6 transformants per microgram plasmid DNA. Although most variables tested, including transformation reagents and competent cell preparation, had little overall effect on transformation efficiency, the efficiency was increased three-fold by maintaining transformed cells at 4°C for 12-18 h prior to plating. Comparison of single and double transformation frequencies showed that processing cells to make them competent for transformation results in a higher frequency of double transformants than if these are independent events. We interpret this to mean that that, similar to *S. cerevisiae*, a sub-population of cells is exceptionally competent for transformation. This is useful because with a frequency of double transformants five-fold lower than that of single transformants, it is practical to indirectly monitor introduction of unmarked genes into a background which may have a limited number of selectable reporter markers available using co-transformation with a single selectable marker plasmid followed by screening.

Unlike in *S. cerevisiae,* where HR is the dominant DNA repair pathway and even 50 bp is sufficient to support efficient targeted integration [23], NHEJ is the dominant form of repair in *Y. lipolytica* and targeted integrations are a small or undetectable fraction of events. Two recent reports examined the effect of deleting genes involved in NHEJ on the frequency of targeted integration at the *TRP1* and *HIS3* [13], and the *ADE2* [26] loci in *Y. lipolytica.* In our study we examined two additional loci for effects of flanking sequence on HR versus NHEJ. Our study together with the previous studies indicates that 500bp of flanking homology only weakly supports HR while longer lengths of 750 and 1000 bp significantly increase the proportion of HR. (Kretzschmar, Otto et al. 2013; Verbeke, Beopoulos et al. 2013). Disruption of NHEJ in all cases significantly increased HR Independent of length. At the *LEU2* locus we showed that the *ku70*□ phenotype conferred a 2.5 fold increase in the fraction of HR-mediated events. Industrial processes can involve elevated temperatures. We found that in spite of roles of Ku70 in stress resistance, heat sensitivity was not significantly affected by the *ku70Δ* mutation.

This study describes an improved protocol for *Y. lipolytica* transformation and a strain modification which increases the proportion of targeted events modifications. As *ku70*Δ cells behave similarly to wt in terms of transformation efficiency and growth, this strain could prove useful in the bioengineering of *Y. lipolytica*. However, cautionary notes also emerged from the current study together with previous studies: First, there are locus-specific differences in HR-mediated integration, and second, homology sequences greater than 500bp may be necessary to significantly increase targeting. Finally there is no assurance that disruption of NHEJ reduces sequence length dependence sufficiently to make PCR-generated fragment knockout strategies feasible (Verbeke, Beopoulos et al. 2013). Thus, additional improvements are important to consider moving forward. In the case of animal cells, which similar to *Y. lipolytica,* have low frequencies of HR, a successful approach was to insert a gene, which can be selected against, outside of flanking homologies so that non-HR isolates can be eliminated (Mansour, Thomas et al. 1988).

With the push for alternative fuels, the need for increased food production, and other avenues to which *Y. lipolytica* can contribute, it is clear that it will be an organism that will be of even greater interest in the future. This work optimizing transformation efficiency and examining the effect of flanking sequence on targeting adds to the knowledge required for its molecular manipulation.

CHAPTER 3: Sequence Assembly of *Yarrowia lipolytica* Strain W29/CLIB89 Shows Transposable Element Diversity

Introduction

The oleaginous yeast *Yarrowia lipolytica* is an industrial model organism for production of biosustainable hydrocarbon-based chemicals (Nicaud 2012; Tai and Stephanopoulos 2013; Pomraning, Wei et al. 2015; Sun and Alper 2015; Zhu and Jackson 2015; Kerkhoven, Pomraning et al. 2016). *Y. lipolytica* is one of the most divergent of the characterized Hemiascomycetes (Dujon, Sherman et al. 2004). Despite a genome almost twice the size of *Saccharomyces cerevisiae*, *Y. lipolytica* is not thought to have undergone whole genome duplication (Gaillardin, Mekouar et al. 2013). In addition, *Y. lipolytica* has more traits in common with metazoan cells than other characterized yeasts. These include dispersed 5S genes, signal-recognition-particle type 7SL RNA sequence, and a greater fraction of the genome composed of introns and intergenic sequences (Dujon, Sherman et al. 2004; Gaillardin, Mekouar et al. 2013). The *Y. lipolytica* genome also contains representatives of diverse classes of transposable elements, including remnants of a DNA transposon (Neuveglise, Chalvet et al. 2005), long-terminal repeat (LTR) (Neuveglise, Feldmann et al. 2002) and non-LTR *L*ong *IN*terspersed *E*lement (LINE) (Casaregola, Neuveglise et al. 2002) retrotransposons (Casaregola and Barth 2013). Finally, unlike the more widely-studied respiro-fermentative *S. cerevisiae*, *Y. lipolytica* is an obligate aerobe. It metabolizes a wide range of carbon substrates including lipids, paraffins, oils, glycerol, and acetate and is capable of accumulating a high percentage of cell weight in lipid (Beopoulos, Nicaud et al. 2011; Tai and Stephanopoulos 2013; Blazeck, Hill et al. 2014). This metabolism has recently been tuned for production of hydrocarbon chemicals.

Availability of an annotated, complete genome assembly is a significant advantage for the study of any organism. The current *Y. lipolytica* genomic reference sequence, YALI0, is that of *Y. lipolytica* strain E150/CLIB122 (hereafter CLIB122) (Barth and Gaillardin 1996; Dujon, Sherman et al. 2004; Gaillardin, Mekouar et al. 2013) (http://www.ncbi.nlm.nih.gov/genome/genomes/194). The YALI0 assembly features the six chromosomes that have been reduced to thirteen contigs and genes that have been extensively annotated [reviewed (Gaillardin, Mekouar et al. 2013)]. CLIB122 was derived from a cross between isolates from a Paris sewer (W29/CLIB89, hereafter CLIB89) and an American corn processing plant (CBS6124-2) (Wolf 1996). Some current strains of industrial interest, including PO1f (Nicaud, Madzak et al. 2002), were derived directly from CLIB89 (Gaillardin, Mekouar et al. 2013; Beopoulos, Verbeke et al. 2014; Blazeck, Hill et al. 2014). Draft reference genomes of *Y. lipolytica* PO1f of 348 contigs (Liu and Alper 2014) and CLIB89 of 369 contigs (Pomraning and Baker 2015) have recently been assembled by alignment with the CLIB122 assembly . However, a complete and independent assembly of strain CLIB89 has been lacking.

We report here the *de novo* assembly and annotation of the *Y. lipolytica* strain CLIB89 genome. Illumina and PacBio sequencing enabled a hybrid assembly of single contigs for chromosomes A-F and mitochondrial chromosome M. Irys long-range genome mapping was utilized to identify extensions of rDNA repeats on the left ends of chromosomes A, C, and F and the right end of chromosome B. Complete sequences of key genetic markers, *URA3* and *LEU2,* and one copy of rDNA sequence not represented in the CLIB122 annotation, were determined. A potentially active copy of Tyl3, a Ty3/Gypsy-like long terminal repeat (LTR) retrotransposon was discovered. Comparison of the sequence of the CLIB89 genome to related strain CLIB122 revealed unexpected differences in numbers and types of transposable elements.

Results and Discussion

Assembly and annotation

Genome assembly

The *Y. lipolytica* genome sequence was determined by HiSeq 2500 (Illumina Inc.) and PacBio RS II (Pacific Biosciences) high-throughput sequencing coupled to a hybrid assembly pipeline (Materials and Methods, Table 3.1, and S1 Text). First, overlapping short, high-quality Illumina HiSeq 2500 sequencing reads were merged into contigs; second, long PacBio reads were used to traverse retrotransposons and bridge the HiSeq contigs, and third, junctions were further refined by aligning with high-quality Illumina reads. PCR was used to confirm key contig junctions (S1 Table). In the next phase, the Irys long-range genome mapping system (BioNano Genomics Inc.) was used to evaluate the integrity of the Illumina-PacBio hybrid assembly, estimate the extent of unassembled sequence in telomeric regions, and localize rDNA repeats (Materials and Methods, Fig 3.1, Table 3.2). The CLIB89 genome assembly was designated YALI1 to distinguish it from the previous CLIB122 YALI0 assembly (previously http://www.genolevures.org/index.html#; CLIB122 YALI0 is now maintained at http://gryc.inra.fr and at http://www.ncbi.nlm.nih.gov/genome/genomes/194) (Dujon, Sherman et al. 2004). Initial comparison of CLIB89 YALI1 and CLIB122 YALI0 assemblies showed that they were similar in both individual chromosomal lengths as well as in total chromosome lengths (Table 3.3). The total length is also similar to that of the draft sequences of the CLIB89-derived PO1f strain of 19,922,824 bp (Liu and Alper 2014) and to that of the CLIB89 20.3 Mb draft sequence (Pomraning and Baker 2015). Both of those were assembled by alignment to the CLIB122 YALI0 reference sequence.

Table 3.1. CLIB89 YALI1 sequence read statistics

Dataset	**Platform**	**Reads**	**Read length**	**Average coverage**
YL97B	SR HiSeq 2000	14,951,623	97	69
YL110	PR HiSeq 2500	389,608,406	110	2041
YLP13	PacBio RS II	157,966	3362	25
YLP14	PacBio RS II	253,645	5642	68

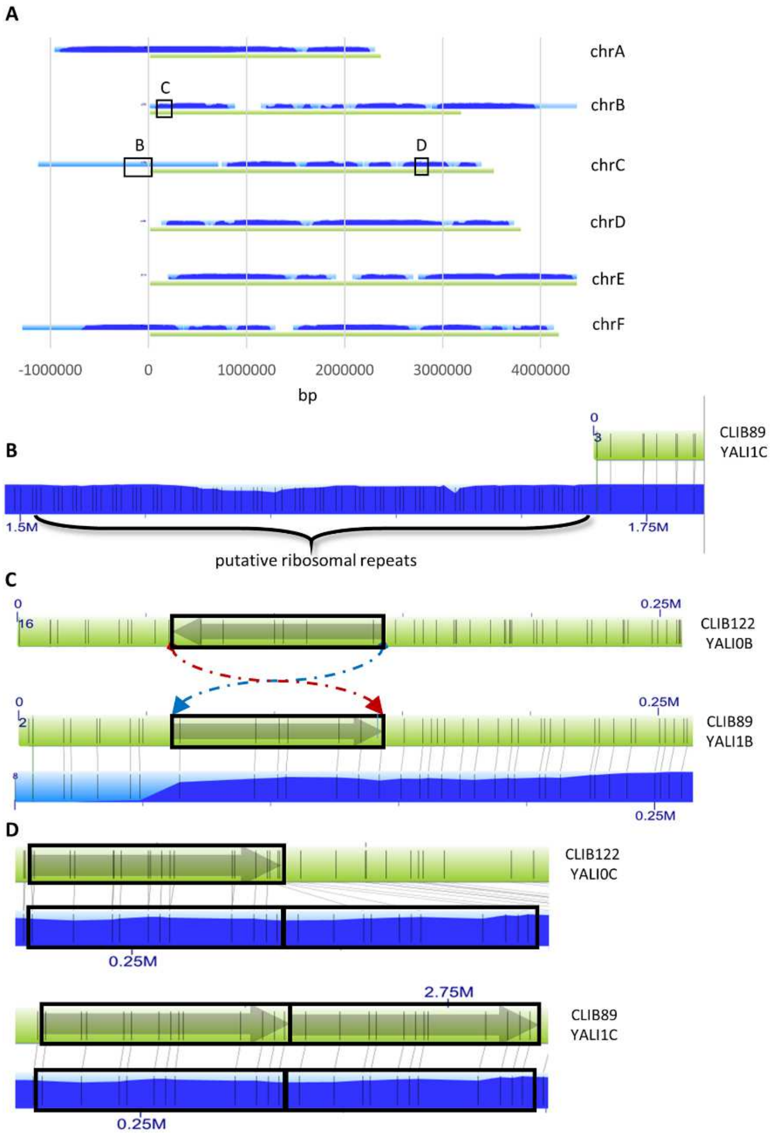


Fig 3.1. BioNano Irys long-range mapping of CLIB89 YALI1 and comparison to CLIB122 YALI0. (**A**) Irys molecules assembled into contigs (coverage indicated as light

and dark for lesser and greater coverage, respectively) aligned with the six CLIB89 YALI1 chromosomes (green) show extensions in four chromosome terminal regions. (**B**) Chromosomal extensions show a repeated pattern of Nt.BspQI nickase sites (vertical grey lines), consistent with tandem, copies of ribosomal (r) RNA-coding sequence. (**C**) and (**D**) Alignment of map of Nt.BspQI sites in the CLIB122 YALI0 (upper green bar) and CLIB89 YALI1 assembly (lower green bar) with the physical Nt.BspQI map generated by Irys technology (blue bar) shows they differ by a 71-kb inversion on chromosome B (**C**) and a 54-kb repeat on chromosome C in CLIB89 YALI1 (**D**).

Table 3.2. Irys CLIB89 YALI1 assembly

Number of consensus genome scaffolds	31
Consensus genome scaffold size (Mb)	25.2
Number of molecules mapped	621,169
Mapped molecule quantity (Mb)	40,798.6
Mapped average size (kb)	269
Average depth of molecule coverage	71.3
Average label density (per 100 kb)	14.4

Table 3.3. Chromosome assembly lengths

Chromosome	**CLIB122 (nts)**	**CLIB89 (nts)**
A	2,303,261	2,257,857
B	3,066,374	3,044,971
C	3,272,609	3,366,276
D	3,633,272	3,629,463
E	4,224,103	4,198,534
F	4,003,362	4,002,965
Total	**20,502,981**	**20,500,066**
M	47,916	47,926

Genome annotation

Fungal genomes differ from some other metazoan genomes in having a high density of coding sequences, low frequency of introns, and in some in codon usage. These differences combined with the medical and agricultural importance of fungi have motivated development of fungal specific analysis strategies. In order to preserve any significant differences between the genomes, the CLIB89 assembly was performed independently of the previous CLIB122 assembly, and identified significant differences from CLIB122. The existence of a previously annotated genome for *Y. lipolytica* CLIB122 was an important asset in this project. This reference genome was used for comparison to identified genes and gene-naming. However, the CLIB89 assembly is the first in which each chromosome is represented by a single contig.

The CLIB89 genome was analyzed and annotated using a combination of parallel pipelines: 1) a customized in-house to search for sequences present in the NCBI *Y. lipolytica* database (http://www.ncbi.nlm.nih.gov/); 2) Yeast Genome Annotation Pipeline (YGAP) (Proux-Wera, Armisen et al. 2012); and 3) SnowyOwl fungal genome analysis (Reid, O'Toole et al. 2014) (Materials and Methods, Fig 3.2). The results of this analysis are summarized in Fig 3.3 and Table 3.4. Subsequent to this primary analysis, multiple comparisons were made between CLIB89 YALI1 and CLIB122 YALI0 genomes**.**

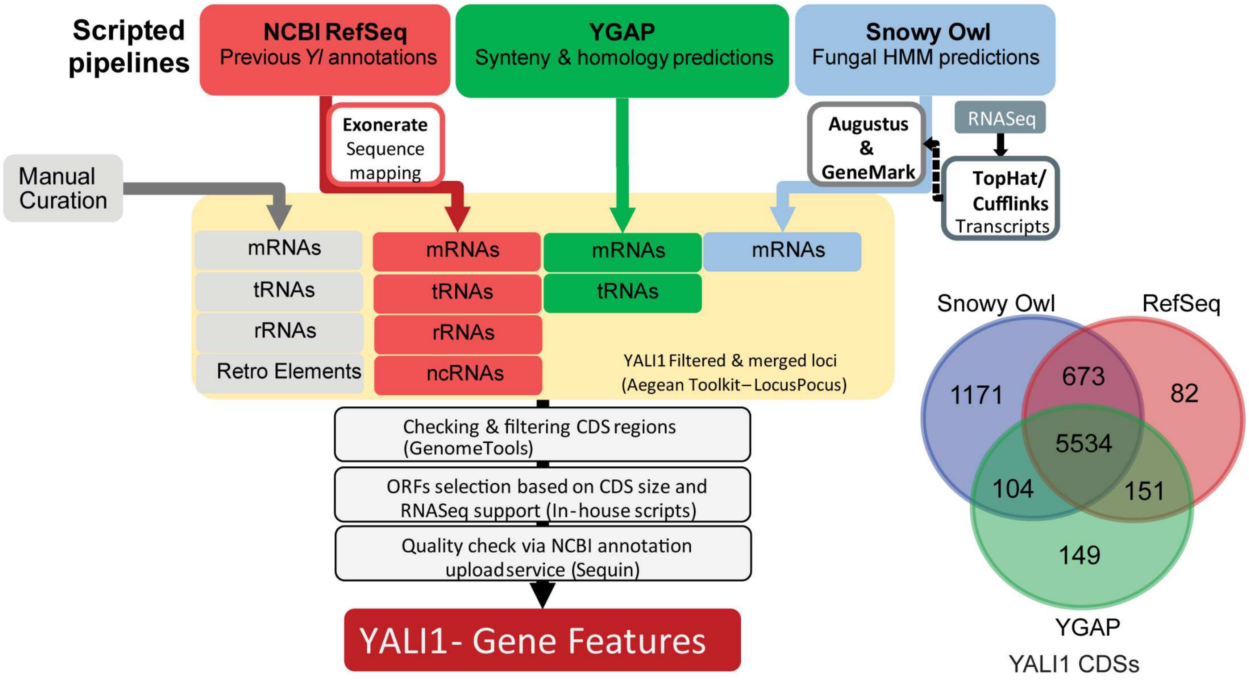


Fig 3.2. *Y. lipolytica* CLIB89 YALI1 *annotation pipeline.* YALI1 Annotations were derived from a combination of three automated annotation pipelines and a set of manual Blast searches. The three pipelines consist of mapping existing *Yl* annotations from CLIB122 YALI0 (NCBI RefSeq) to the CLIB89 YALI1 sequence; synteny and homology predictions (YGAP); and fungal HMM predictions (Snowy Owl). Loci of identified features were merged, checked for consistency, selected for CDS based on size and RNA-Seq support, and vetted through NCBI's Sequin upload service to produce the final set of gene features. Contributions and agreements for CDS features from the three automated pipelines are shown in the Venn diagram to the lower right.

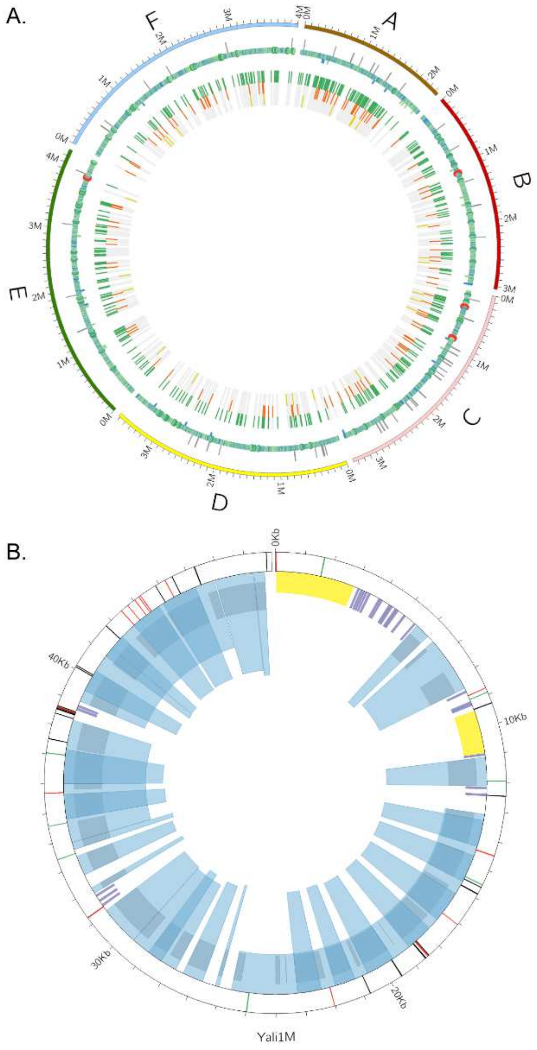


Fig 3.3. CIRCOS overview of YALI1 gene features. (A) Chromosomal genes. Outer ring, chromosomes. First mapping track [RNA polymerase II (POL2)-transcribed genes]: LINE retroelements (outward light grey posts), overlapping genes on both strands (inward blue posts), POL2 less than 1 kb (blue rings), POL2 between 1 kb and 5 kb (light green rings), POL2 between 5 kb and 10 kb (green rings), and POL2 > 10 kb (red rings). The next inner track (POL3 and POL2 ncRNA genes), tRNA (green), rRNA (orange), and ncRNA genes (dark yellow). **(B)** Mitochondrial genes. Transcripts from exons (longer spanning light blue wedges); transcripts from introns (narrower and taller overlapping wedges); CDS (gray); tDNA (purple); and rDNA (yellow). Outer track shows variants comparison with the CLIB122 assembly （http://www.ncbi.nlm.nih.gov/nuccore/NC_002659.1): mismatches (black posts); insertions (green) and deletions (red) relative to the CLIB122 assembly.

Overall there is high consistency between the CLIB89 YALI1 genome reported here and the previous CLIB122 YALI0 assembly. The NCBI annotated data from *Y. lipolytica* is largely based on the CLIB122 assembly (Dujon, Sherman et al. 2004) (http://www.ncbi.nlm.nih.gov/genome/genomes/194), but also includes other *Y. lipolytica* sequences. We refer here to these annotated data as the RefSeq Yl (database) although it is not formally a single curated RefSeq genome sequence. From the protein coding sequences in CLIB122 RefSeq Yl, 31 had no sequence matches to CLIB89 YALI1, while 6,582 matches were found in CLIB89 YALI1 using Exonerate software (Slater and Birney 2005). After filtering and combining with other annotation pipelines, the RefSeq mapping annotated 6,440 CDS out of the combined total of 7864 CDS, with 82 CDS unique to this pipeline (Fig 3.2).

For additional gene discovery, CLIB89 YALI1 was analyzed using YGAP software (Proux-Wera, Armisen et al. 2012). YGAP was developed for fungal genomes and exploits the existence of a large number of gene sequences ("pillars") conserved among fungal species and maintained in the Yeast Gene Order Browser (YGOB) database (Byrne and Wolfe 2005) as well as the syntenic arrangement of coding regions among a large number of fungi.

Availability of RNA-seq data from a parallel study in our laboratory enabled SnowyOwl, a fungal Hidden-Markov model (HMM) gene predictor approach that uses transcriptome data for model validation (Reid, O'Toole et al. 2014). The SnowyOwl pipeline leveraged a combination of RNA-Seq and gene homology searches for *ab initio* gene prediction.

Table 3.4. CLIB89/CLIB122 gene content

	CLIB122[2]	CLIB89[2]	Unique CLIB122	Identified In CLIB89[5]
mRNA[1,2]	6472 CDS	7864 CDS +118 pseudo	31	1428 CDS +88 pseudo
tDNA YALI [A-F][3]	510 +9 pseudo	509 +9 pseudo	5	5
tDNA YALI [M]	26	24	3	1
5S rDNA [A-F]	117	111 +3+4* pseudo	9	9
Other ncDNA[4]	12+13 Ruf70	15+13 Ruf70		3
Total features	7150 +9 pseudo	8536 +134 pseudo	48	1534

[1] RefSeq; reported as CLIB122 “proteins”; CLIB122http://www.ncbi.nlm.nih.gov/genome/genomes/194); identified in CLIB89 are CDS identified in the pipelines as described in Materials and Methods; parallel analysis was not performed on CLIB122

[2] Pseudo refers to pseudogenes that resemble original genes, but have interrupted coding sequences

[3] Eight tDNA-like sequences present in both strains but not previously reported for CLIB122 are included and not counted as unique to CLIB89

[4] Thirteen Ruf70 present in both assemblies but not annotated in CLIB122

[5] Annotated in CLIB89 YALI1, but not in CLIB122 YALI0

* Pseudo rRNA too short to include in NCBI annotation

Final merged set of annotated genes

The three sets of gene loci and a handful of manually curated loci were merged into one final set based on common coordinates in the CLIB89 YALI1 assembly. For each protein-coding gene, an optimal ORF configuration was defined based on the best agreement among the three pipelines. The locus ID numbering in YALI0 was revised in YALI1 to accommodate additional sequences assembled at the ends of chromosomes. Common names based on *S. cerevisiae* genes and utilized in the YALI0 annotation were retained. Sequences not previously identified as genes were designated following standard nomenclature for Hemiascomycete yeasts and consistent with CLIB122 nomenclature. [Four capital letters refer to the genus and species, one digit to the assembly, A to F and M to the chromosomes and five digits to the coding regions followed by g for proteins, t for repeated sequences, r for non-coding sequences, and s for cis-acting features (Dujon, Sherman et al. 2004; Durrens and Sherman 2005).] The CLIB89 YALI1 coding sequences mapping to CLIB122 YALI0 and annotations, together with YALI1 designations for mRNA, noncoding (nc) and retrotransposon sequences and corresponding YALI0 designations are shown in the S2 Table. Thirty-one genes were identified as unique to CLIB89 (S5 Table). The annotated assembly is available through a browser interface at the Institute for Genomics and Bioinformatics (https://www.igb.uci.edu/), the Sandmeyer laboratory (http://sbsngsserver.biochem.uci.edu/jbrowse/index.html), and has been uploaded to the NCBI website (http://www.ncbi.nlm.nih.gov/). Details of this process are provided in Materials and Methods. Annotations of CLIB89 features are available for uploading and viewing in a genome browser (S3 and S4 Tables) (http://gmod.org/wiki/GFF3).

Genomic features

Chromosomal terminal sequences and rDNA

Four chromosomes were previously reported to contain terminal clusters of rDNA sequence (van Heerikhuizen, Ykema et al. 1985; Casaregola, Feynerol et al. 1997; Gaillardin, Mekouar et al. 2013). Nonetheless, the complete rDNA 35S sequence was not determined in either the original CLIB89 YALI1 or CLIB122 YALI0 assemblies. BLAST analysis utilizing *S. cerevisiae RDN18-1* sequence to query CLIB122 YALI0 sequence identified a small segment of the rDNA sequence near the end of chromosome F with rDNA similarity. Primers (S1 Table) complementary to sequence near the end of CLIB89 chromosome F and to sequence containing the CLIB122 rDNA segment were used to amplify the intervening DNA from CLIB89 DNA. This sequence contained rDNA sequence thus positioning one copy relative to CLIB89 YALI1 sequence. An overlapping sequence was recovered from high fold coverage unmapped reads from the CLIB89 YALI1 sequence and used to design primers that enabled amplification of an almost full-length copy of the rDNA repeat and adjacent sequences. The products of these PCR reactions were analyzed by Sanger sequencing. BLAST analysis of the assembled sequence showed sequence consistent with the order External Transcribed Spacer (ETS), 18S rRNA, Internal Transcribed Spacer 1 (ITS), 5.8S, ITS 2, 25S rRNA, and flanking NTS sequences (S2 Text). Previous analysis of the *Y. lipolytica* rDNA locus showed that the NTS may be present in forms of different lengths (van Heerikhuizen, Ykema et al. 1985). On average, the fold coverage of

Illumina sequences was approximately 100X greater than that for uniquely-aligned reads.

Informed by the complete rDNA sequence, a virtual digest with Irys nickase Nt.BspQI was performed to determine the rDNA pattern predicted to be generated by this enzyme. The Nt.BspQI predicted rDNA pattern was compared to the Nt.BspQI pattern of telomeric regions where Irys *Y. lipolytica* molecules extended beyond the assembly. This comparison identified predicted rDNA pattern repeats localized to the left ends of chromosomes A, C, and F and the right end of chromosome B (Fig 3.1B). This result corroborates previous proposals that the rDNA of *Y. lipolytica* is comprised of distributed telomere-proximal clusters, and for the first time provides a molecular map of the number and location of these clusters.

In most eukaryotes, maintenance of chromosomal ends is through telomerase 3' terminal extension using a short guide RNA. The updated *Y. lipolytica* telomeric sequence (5'GGGTTAGTCA3') (Gaillardin, Mekouar et al. 2013) matched highly-repeated, unassembled reads in our CLIB89 sequences. Due to the repeated nature of these sequences and the absence of Nt.BspQI nickase sites, they could not be assembled into the complete CLIB89 YALI1 genomic sequence.

Small non-coding RNA genes

RNA Polymerase III-transcribed genes (RNAP3 genes) in *Y. lipolytica* have interesting features that distinguish them from RNAP3 genes in other yeasts (Dujon, Sherman et al. 2004; Marck, Kachouri-Lafond et al. 2006; Acker, Ozanne et al. 2008; Neuveglise, Gaillardin et al. 2013). First, there are about twice as many; second, the 5S genes are dispersed throughout the genome, rather than clustered within rDNA repeats; and third, dimeric RNAP3 genes are abundant. RNAP3 genes are described in Table 3.4 and S6 Table. Dimeric genes are characterized by RNAP3 gene-coding sequences separated by only 5 to 26 nts with the upstream member followed by an abbreviated RNAP3 terminator tract of T's.

From RefSeq Yl mapping, 510 CLIB122 YALI0 tDNA sequences from chromosomes A-F were identified in CLIB89 YALI1. In addition, nine other tDNA-like sequences that did not pass tRNAscan-SE, but had high similarity to tDNA from RefSeq Yl and were not annotated in CLIB122 YALI0, were identified as allelic in the two genomes (Table 3.4, S6 Table). These additional nine genes were classified as tDNA pseudogenes. RNAP3 genes have multiple interactions with retroelements and we speculate that these pseudogenes arose as reverse transcript cDNAs of tRNAs that were integrated into the genome.

A significant fraction of RNAP3 genes are present in multiple copies. In *S. cerevisiae* the 5S rDNA occurs within spacers between tandem rDNA repeats. However, in *Y. lipolytica* and most metazoan species, the 5S genes are dispersed outside the rDNA repeats (van Heerikhuizen, Ykema et al. 1985; Clare, Davidow et al. 1986; Acker, Ozanne et al. 2008). One hundred and eleven 5S genes and seven pseudo 5S genes were identified in CLIB89 of which 103 were perfectly matched to CLIB89 5S genes (Table 3.4, S6 Table, YALI1 5S gene summary). Of the seven pseudo 5S genes, four were below the length threshold to be submitted to NCBI.

Analysis of the CLIB89 YALI1 genome using Exonerate/BLAST identified single representatives of the RNAP3 *SNR52*, *SCR1*, *SNR6*, and *RPR1* genes and thirteen copies of *RUF70,* as previously annotated in CLIB122 YALI0 (Marck, Kachouri-Lafond et al. 2006; Acker, Ozanne et al. 2008; Neuveglise, Gaillardin et al. 2013). Similar to CLIB122 YALI0, copies of Ruf70 were downstream of tDNA^{Trp}(CCA) in CLIB89 YALI1. In addition, genes encoding RNAP2-transcribed processing RNAs U1-U5 and U7 RNAs were identified.

Differences between CLIB89 and CLIB122 genomes

Structural variation

CLIB89 YALI1 and CLIB122 YALI0 genomes were globally compared to themselves and to each other by alignment of the assemblies using a dot matrix program, MUMmer (http://mummer.sourceforge.net/) (S1 Fig). Self-alignment highlighted, as expected, repeated sequences throughout the two genomes. Many of these were 5-6 kb or 300-500 bp in length, consistent with the sizes of full-length or truncated LINE elements and LTR retrotransposons or solo LTRs. However, consistent with the Irys map (Fig 3.1C) comparison of the CLIB89 and CLIB122 assemblies revealed a 71-kb sequence in CLIB89 YALI1 chromosome B that was inverted relative to the same sequence in the CLIB122 YALI0 assembly and the draft CLIB89-related assemblies, PO1f and W29 (Souciet, Aigle et al. 2000; Dujon, Sherman et al. 2004; Liu and Alper 2014). PCR was performed across the upstream and downstream junctions of the inversion using primer pairs JY5118/JY5119 and JY5120/JY5121 respectively (S2 Fig). These reactions generated products consistent with the predicted sizes from the CLIB89 YALI1 assembly of 752 bp and 763 bp, respectively. Products with an identical migration pattern were generated when PO1f genomic DNA was used as a template.

Furthermore, although a negative result, primer pairs JY5118/JY5121 and JY5119/5420, failed to generate a product, contrary to what was predicted in the CLIB122 YALI0 assembly. This result is consistent with the close relationship of CLIB89 and PO1f.

MUMmer analysis also revealed a striking ~115-kb region in CLIB89 YALI1, but not CLIB122 YALI0. Examination of the DNA sequence in this region showed that, consistent with the Irys map (Fig 3.1D), a 54-kb sequence was repeated with three 6.5-kb Ylli LINE copies: flanking upstream and downstream and separating the 54-kb repeats. In order to validate this structural variant, PCR was performed using primers JY5124/JY5125 (S1 Table, S2 Fig), complementary to sequences at the downstream and upstream junctions of the 54-kb sequence with the central Ylli sequence. PCR generated a product consistent with the size predicted for an amplicon containing Ylli sequence bounded by head and tail ends of flanking copies of the 54-kb sequence. In the CLIB122 YALI0 assembly, the 54-kb block occurs as a single copy flanked by Ylli sequence direct repeats. These differences between the two assemblies could have been generated by unequal crossing over between the first and second Ylli elements. Local amplification of gene copy number such as this might have advantages for the host. In this case, it is difficult to speculate. There were a total of 29 features including 27 CDS among them genes implicated in stress resistance.

IrysView software was used to align the physical CLIB89 Irys Nt.BspQI molecules and virtual sequence based CLIB89 YALI1 Nt.BspQI patterns. This comparison showed consistency between the physical molecular and sequence maps, thereby confirming the orientation of a 71 kb sequence and the existence of the 54-kb direct repeat flanked by Ylli element fragments in the CLIB89 YALI1 sequence (panels C and D Fig 3.1).

The CLIB89 YALI1 sequence was examined for the region of chromosome A encompassing four protein-coding genes found in CLIB122, but reported to be absent in the latest PO1f draft assembly (CLIB122: 196442 – 215157)(Liu and Alper 2014). This sequence in CLIB122 YALI0 included YALI0A01562 and YALI0A01602, genes encoding proteins with weak similarity to SMC5/6 proteins involved in DNA repair and recombination, leading to speculation that this deficiency contributes to the relatively low ratio of homologous recombination (HR) to non-homologous end joining (NHEJ) observed in *Y. lipolytica* (Liu and Alper 2014). Similarly, our assembly showed that this segment is absent from chromosome A in CLIB89. However, BLAST search of the YALI1 sequence revealed a 93% match to the sequences of YALI0A01562 on chromosome E (YALI1E20467g) and a 95% match to YALI0A01602g on chromosome F (YALI1F03604).

Protein-coding differences

After annotation, CLIB89 YALI1 and CLIB122 YALI0 genomes were compared, revealing unique features in each (S2, S5 and S6 Tables). Differences between CLIB89 YALI1 and YALI0 transposons are discussed below. CLIB89 is a wild-type strain so that *URA3* and *LEU2* sequences are represented. However, *URA3* and *LEU2* (Kretzschmar, Otto et al. ; Beopoulos, Verbeke et al. ; Blazeck, Hill et al.) genes were deleted and disrupted, respectively, in CLIB122 to enable the intact genes to be used as selectable markers. *URA3* encodes orotidine-5’-phosphate decarboxylase, a central enzyme in the uracil biosynthetic pathway. *LEU2*, encodes beta isopropylmalate dehydrogenase, which is critical for leucine biosynthesis and also an important genetic marker (Clare, Davidow et al. 1986; Gaillardin, Mekouar et al. 2013; Tai and Stephanopoulos ; Blazeck, Hill et al.). In CLIB122 the *LEU2* ORF, YALI0C00407g, is disrupted by insertion of the *S. cerevisiae SUC2* gene for invertase, which allows for growth on sucrose (Nicaud, Fabre et al. 1989). *Y. lipolytica* is thought to be heterothallic. Therefore, rather than switching expression of a mating-type locus, strains themselves are of alternative Mating Types, A and B. CLIB89 possesses the Mating-Type A protein and CLIB122 possesses the Mating-Type B protein, consistent with the previously reported mating types (Weber and Kurischko 1989). Twelve additional genes were identified in terminal regions of the chromosomes in the CLIB89 assembly that were not present in the CLIB122 assembly (Table S5). The simplest explanation is that rather than constituting an actual difference between the strains, this region was incomplete in the CLIB122 YALI0 assembly. These regions contained: one heat-shock gene on chromosome A; a block of seven genes at the end of chromosome D, and four genes within the first 32 kb of chromosome E. This interpretation is supported by sequence reads related to argininosuccinate synthase that we report to be encoded in the terminal end of chromosome D in CLIB89 YALI1, that were previously reported collected in the CLIB122 sequencing project, but not mapped to any chromosomal scaffold (Gaillardin, Mekouar et al. 2013).

Comparison of genes identified in CLIB89 YALI1 with those reported in CLIB122 YALI0 also identified several that were present in CLIB122 YALI0, but not identified in CLIB89 YALI1. Differences in the mating-type proteins were among these, as expected. In addition, there were multiple differences in transposon sequences of class I (RNA) and class II (DNA) elements further detailed below. In addition, we report 1428 CDS and 88 pseudogenes (Table 3.4) in CLIB89 YALI1 sequence. The majority of the 1428 CDS were discovered in the SnowyOwl pipeline and lacked transcript or identifiable protein domain support. In addition, compared to the CDS with those properties, a significant fraction of these CDS were relatively short, although longer than 100 codons.

Transposable elements in CLIB89

Y. lipolytica displays striking diversity in transposon composition with relics of a Class II (DNA) element and both LTR and non-LTR class I (RNA) elements (Table 3.5 and S7 Table). Full-length copies of LTR retroelements can undergo recombination between the LTRs resulting in deletion of the internal domain and one LTR copy, generating a so-called solo LTR. Despite fundamental differences in transposition mechanisms, transposable elements (TE) have common distinguishing features. For example, they are typically present in multiple copies per genome. Ultimately both classes are mobilized by transposases/integrases and terminate in a conserved inverted repeat recognized by those enzymes. TG…CA represents the minimal virtually universally conserved terminal inverted repeat, although individual elements exhibit a range of inverted repeat lengths. Insertions are initiated by strand transfer. Because these reactions occur across a DNA helix, reacting positions on the two strands are offset; repair of the single-stranded extensions caused by the offset generates short target-site duplications that flank outside transposon ends (Craig 2015). We first searched for Ty1-Copialike and Ty3-Gypsylike conserved reverse transcriptase-coding sequences, and other TE sequences identified in CLIB122 YALI0. However, during identification of allelic tDNAs, we also identified a number of insertion polymorphisms which were characterized by the properties described above that are conserved among solo LTRs. Because we failed to identify a retroelement full-length copy, we designated these as LTRyl7, 8, 9. However, this did not represent an exhaustive search of the CLIB89 genome for novel LTRs.

Table 3.5. Families of transposable elements in CLIB89 and CLIB122

Class of Transposable Element	**CLIB89**	**CLIB122**	**Allelic**
Ylli - L1 non LTR retroelement			
# of Full length (6494 bp)	17	10	7
# of Partial (~3800 bp)	4	2	0
Ylt1 - Ty3/Gypsy retroelement			
# of Full length (9453 bp)	0	10	n/a
# of Solo LTRs (~715 bp)	0	17	n/a
Tyl6 - Ty3/Gypsy retroelement			
# of Full length (5.104 kb)	0	1	n/a
# of solo LTRs (276 bp)	0	0	n/a
Tyl3 - Ty3/Gypsy retroelement			
# of Full length (5973 bp)	1	0	0
# of Solo LTRs (244 bp)	4	4	3
Mutyl - Mutator-like DNA transposon			
# of Full length (7413 bp)	0	5	0
# of solo MudrA (3537 bp)	1	1	1
# of solo MudrB (1380 bp)	0	0	n/a
# of solo MudrA like (2628 bp)	1	1	1
# of solo MudrB like (1379 bp)	1	1	1
LTRyl1 element (Putative)			
# of Full length	0	0	n/a
# of Solo LTRs (~278 bp)	54	30	30
LTRyl7 element (Putative)			
# of Full length	0	0	n/a
# of Solo LTRs (~337 bp)	14	17	14
LTRyl8 element (Putative)			
# of Full length	0	0	n/a
# of Solo LTRs (~302 bp)	11	17	9
LTRyl9element (Putative)			
# of Full length	0	0	n/a
# of Solo LTRs (~444 bp)	1	5	1

DNA Class II elements

After discovery of the DNA transposon Mutator (Mu) in maize (Robertson 1978), Mu and MUtator-Like Elements (MULEs) were found in a number of other species (Lisch 2002). The fungal MULE, Hop, in *Fusarium oxysporum* is a well-characterized example (Chalvet, Grimaldi et al. 2003). Hop has 99-bp perfect terminal inverted repeats and generates a 9-bp target site duplication. Transposition is mediated by a Hop-encoded 836-aa transposase related to the bacterial *mudrA*-encoded MURA protein. The first report of any DNA element in *Saccharomycotina* was the MULE Mutyl discovered in *Y. lipolytica* (Neuveglise, Gaillardin et al. 2013). Mutyl elements have imperfect terminal inverted repeats of 22 bp and are flanked by 9-bp target site duplications. Mutyl encodes a transposase, MudrA, and a second protein of unknown function, MudrB, that is not universally present in non-maize active MuDR elements.

Five Mutyl elements were reported in the CLIB122 YALI0 genome, of which four were full-length and one had an 8-bp deletion. There is evidence for at least one recent Mutyl transposition (Neuveglise, Chalvet et al. 2005). In that study, blotting with a MudrB probe failed to identify a complementary sequence in the CLIB89 genome. Given the relatedness of the CLIB122 and CLIB89 strains, this was interpreted to mean that the element might have been acquired by more recent horizontal transmission to the CLIB89 descendant, CLIB122. Consistent with the previous report, MudrB-coding sequence was not found in the CLIB89 YALI1. However, BLAST searches of CLIB89 sequence identified an ORF allelic to CLIB122 YALI0A14971g (YALI1A15017t) the MudrA transposase sequence in the full-length Mutyl element (YALI0A16207) (Table 3.5 and S7 Table). Thus, partial Mutyl sequence is present in the predecessor strain at the same site as in CLIB122 and were likely shared by vertical transmission. Furthermore, both CLIB89 and CLIB122 both contain isolated ORF sequences that encode proteins similar to MudrA (YALI0A02266g/ YALI1A02682g) and MudrB (YALI0C17193g/ YALI1C24526g). However, because these sequences share little nucleotide similarity with MudrA and MudrB, respectively, these are likely only distantly related to the Mutyl family of transposons.

Fotyl represents a second Class II element that has been identified in *Y. lipolytica* (GenBank: CAG33729). This family was discovered as the Fot1 element in *Fusarium oxysporum* (Daboussi, Langin et al. 1992; Daboussi and Capy 2003). A full-length element of this family was discovered in *Y. lipolytica* and dubbed Fotyl. It is a pogo-like element member of the Tc1-Mariner superfamily. A complete copy including terminal inverted repeats was identified on chromosome E and a partial degenerate copy on chromosome A of CLIB122 YALI0 (Gaillardin, Mekouar et al. 2013). These are allelic with sequences in CLIB89 YALI1.

LTR retrotransposons

LTR retrotransposons populate CLIB89 and CLIB122 genomes (Table 3.5, S7 Table). These elements are similar to retroviruses in that the upstream LTR contains the promoter for transcription of genomic RNA and the downstream LTR specifies transcription termination and polyadenylation (Symer and Boeke 2010; Sandmeyer, Patterson et al. 2015). Retrotransposons typically contain two ORFs. The upstream ORF encodes capsid structural and nucleic acid binding domains and the downstream ORF encodes proteinase, reverse transcriptase and integrase. Elements have devised various mechanisms to ensure an excess of structural proteins over catalytic proteins. These include programmed frameshifting or even splicing to join the respective protein sequences. Two superfamilies of LTR retrotransposons are abundantly represented in eukaryotes: Ty1/Copia and Ty3/Gypsy. However, unlike *S. cerevisiae* in which Ty1/Copia elements predominate, no Ty1/Copia elements have been reported in *Y. lipolytica*.

At present, two full-length Ty3/Gypsy LTR retrotransposons, Ylt1 (Schmid-Berger, Schmid et al. 1994) and Tyl6 (Kovalchuk, Senam et al. 2005) have been described for *Y. lipolytica*. Additionally, a partial integrase coding sequence adjacent to an LTR was designated Tyl3, but no full-length sequence has been reported (Casaregola, Neuveglise et al. 2000; Souciet, Aigle et al. 2000; Neuveglise, Feldmann et al. 2002). In the CLIB122 YALI0 assembly, Ylt1 is abundantly represented with 10 full-length and 17 solo LTRs (Schmid-Berger, Schmid et al. 1994) (this study). Tyl6 is represented by one full-length copy and no solo LTRs. Tyl3 is represented by no full-length elements but four solo LTR copies. LTRyl1 is represented by 30 copies of solo LTR sequence (Casaregola, Neuveglise et al. 2000). As noted above, CLIB122 was derived from a cross between French strain CLIB89 and American strain CBS6124-2. Amazingly, CLIB89 YALI1 and CLIB89-derived PO1f completely lack both full-length and solo LTR copies of Ylt1 and Tyl6 based on hybridization, draft assembly and complete sequence analysis (Schmid-Berger, Schmid et al. 1994; Kovalchuk, Senam et al. 2005; Liu and Alper 2014) (this study). However, it has four copies of Tyl3 LTRs three of which are allelic with copies in CLIB122 and 54 copies of LTRyl1 thirty of which are allelic with copies in CLIB122 YALI0. Similar to CLIB122, it lacks full-length versions of the element.

After identification of several novel LTR sequences, we sought to identify novel full-length elements. BLAST analysis of CLIB89 YALI1 sequence used the Ty1/Copia and Ty3/Gypsy conserved reverse transcriptase domains as queries. A truncated *POL* encoding a Ty1-homologous RNaseH was identified adjacent to a copy of LTRyl1. This LTR was not flanked by direct repeats of a target sequence, consistent with it representing the downstream LTR of a truncated element rather than a solo LTR derived from recombination of terminal LTRs. Ty1/Copia and Ty3/Gypsy elements differ in the order of integrase and reverse transcriptase/RNaseH domains. The position of the RNAseH-coding sequence proximal to LTRyl1, the similarity of the sequence to Ty1 RNAseH, and the lack of short direct repeats bordering the LTR argue that the LTRyl1 family is derived from an extinct Ty1/Copia element. This represents the first identification of a Ty1/Copia element in *Y. lipolytica*.

Full-length Tyl3, a Ty3/Gypsy retrotransposon

As described above, tBLASTn analysis was used to search CLIB89 YALI1 coding sequences for those encoding reverse transcriptase, the most highly-conserved retroelement domain. In addition to the Ty1/Copia ORF described above, an ORF with 41% protein identity to the *S. cerevisiae* LTR retrotransposon Ty3 reverse transcriptase domain was identified on chromosome C (S3 Text). A tBLASTn search for the conserved core domain of Ty3 integrase showed a predicted sequence with 62% similarity and a BLASTn search showed a 100% match to the previously-reported integrase carboxyl-terminal sequence of Tyl3 reported in CLIB89 (AL414488, AL414575) (Kovalchuk, Senam et al. 2005). Comparison of the candidate LTR of this element and Tyl3 LTR sequences showed 100% identity. Sequences with similarity to nucleocapsid, and the protease active site confirmed the presence of Gag and Pol-like domains. The two ORFs are flanked by the previously described Tyl3 LTR sequences with 6-bp terminal inverted repeats (TGTAAG/CTTACA) (Fig 3.4A). The outside ends of the LTRs were flanked by 5-bp target site duplications offset by one "T" in the upstream repeat (ATTTTT/ATTTT). Other features of LTR retrotransposons are also present: two nts downstream of the upstream LTR is a sequence complementary to the 3' terminal 14 nts of initiator tRNA^{Met}(CAU), the presumed minus-strand primer. Some patches of initiator tRNA^{Met} complementarity were found in the downstream LTR as well, possibly indicating a bi-partite primer (Gabus, Ficheux et al. 1998). Just upstream of the 3' LTR in Tyl3 and corresponding to the position of the polypurine tract (PPT) plus-strand primer for reverse transcription of LTR retroelements is a sequence of 13 consecutive purines. One full-length copy of this element in addition to four solo LTRs were identified in the CLIB89 genome. We conclude that this element constitutes a full-length Tyl3, a fragment of which was previously reported present in CLIB89 based on partial integrase and LTR sequence (Neuveglise, Feldmann et al. 2002) (Fig 3.4A).

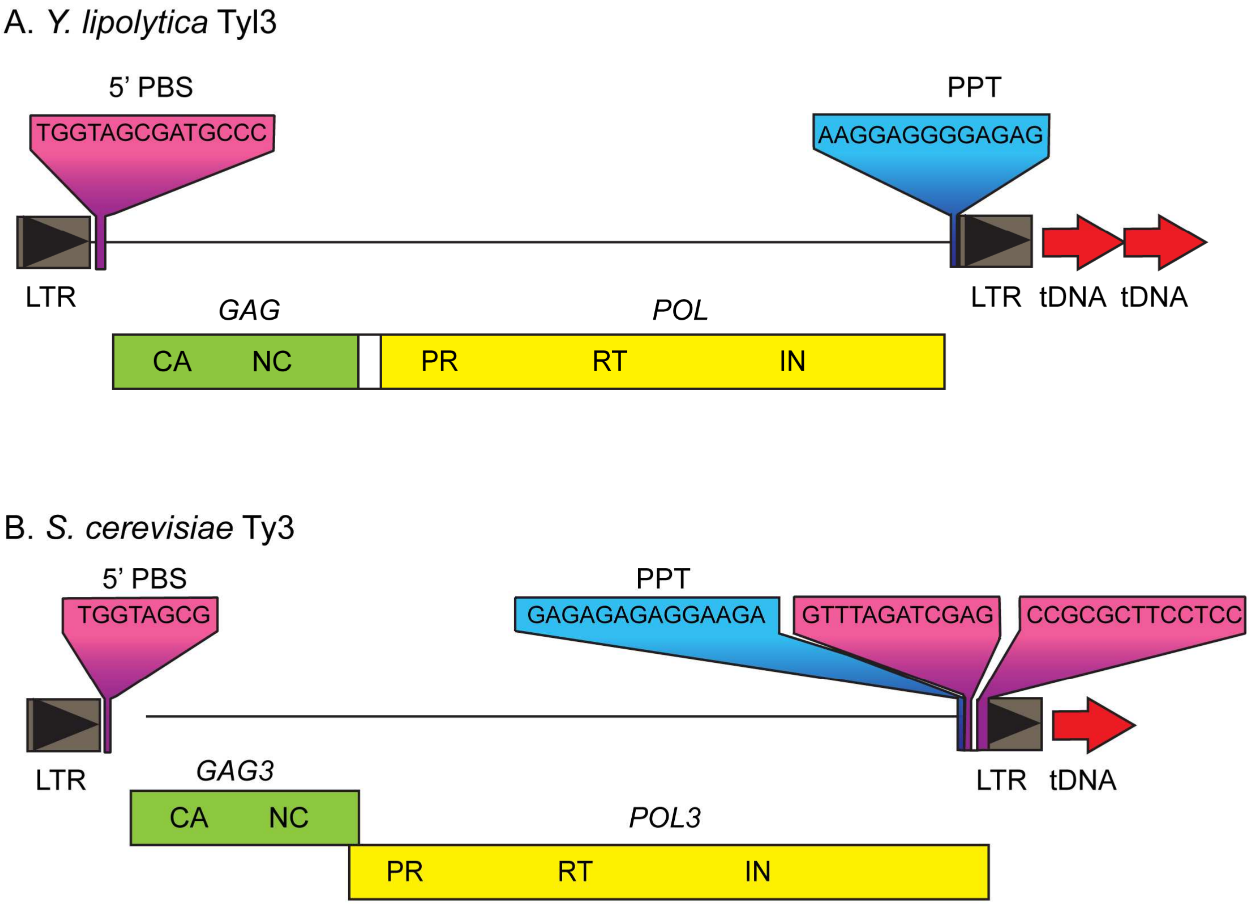


Fig 3.4. The complete Tyl3 *Y. lipolytica* Ty3/Gypsy element. (A) Tyl3 was assembled from CLIB89 sequence. Abbreviations are *GAG* (capsid, CA; and nucleocapsid, NC) and *POL* (protease, PR; reverse transcriptase, RT; and integrase, IN), Solid triangles represent LTRs, *GAG* and *POL* are separated by 324 bp. PBS, primer binding sequence complementary to initiator tRNA^{Met} the presumed primer for minus-strand replication; and PPT, polypurine tract the presumed primer for plus-strand replication. The full-length Tyl3 is adjacent to two tDNA sequences. **(B)** Full-length *S. cerevisiae* Ty3 is shown for comparison. Features are similarly abbreviated as in 4A.

Unlike the vast majority of Ty3/Gypsy elements found in yeasts that contain a frameshift between the *GAG* and *POL* as illustrated by the prototypic element Ty3 (Fig 3.4B), the two Tyl3 ORFs, are separated by 324 bp, which includes four stop codons downstream of the *GAG* stop codon Although this is an unusual configuration, it is not without precedent. The Copia-like element *C. albicans* Tca2 has a stop codon separating the two ORFs (Holton, Goodwin et al. 2001), and an internal promoter has been proposed to explain differential expression of structural and catalytic proteins (Forbes, Nieduszynska et al. 2007). Some viruses express ORFs from internal ribosomal entry sites (Walsh and Mohr 2011) and the foamy retrovirus expresses its *POL* functions from a spliced RNA (Yu, Baldwin et al. 1996). However, scanning the *Y. lipolytica* Tyl3 spacer with software designed to identify viral internal ribosome entry sites (Hong, Wu et al. 2013) and intronic sequences (http://genes.mit.edu/GENSCAN.html)(Burge 1998) failed to identify either.

Inspection of the Tyl3 allelic site sequence in CLIB122 confirmed absence of Tyl3 sequence or any repeat of the 5 bp sequence duplicated at the ends of the full-length Tyl3 in CLIB89. This result suggests that this empty site in CLIB122 was inherited from the American strain CBS6124-2 or that the transposition in CLIB89 was relatively recent followed by loss of the progenitor full-length element.

Members of the Ty1/Copy and Ty3/Gypsy LTR retrotransposon classes display patterns of targeting by integrase to genomic histone marks (chromodomain class) or RNAP3 genes (Malik and Eickbush 1999; Curcio, Lutz et al. 2015; Sandmeyer, Patterson et al. 2015). A subset of Ty3/Gypsy elements, including the eponymous Ty3, exclusively target RNAP3 transcription start sites (Sandmeyer, Patterson et al. 2015).

Because tDNA are transcribed into pre-tRNAs which have an approximately 10-nt pre-sequence, the insertion of a 5-bp target site duplication positions retrotransposon sequence 15 to 17 bp upstream of mature tRNA-coding sequence. Sequences flanking Tyl3 and Tyl3 LTRs were analyzed using tRNAscanSE. Analysis showed that Tyl3 is inserted at the likely transcription start site of tDNA^{Val}, previously identified in CLIB122 YALI0 sequence. In addition, two of the four Tyl3 solo LTRs are positioned at likely transcription start sites of tDNAs. Two others are less closely related and are inserted 5 bp inside the upstream end of a tDNA and at the 5' end of a 5S rRNA gene. A full-length copy of Tyl6 was previously identified close to the transcription initiation site of the tRNA^{Met} gene (Kovalchuk, Senam et al. 2005). Comparison of the integrase domains of these elements showed that they are Ty3/Gypsy elements of the class lacking a chromodomain and closely associated with RNAP3 transcription start sites (Hansen, Chalker et al. 1988; Malik and Eickbush 1999; Kovalchuk, Senam et al. 2005) (this study).

tDNA-associated LTRs

Analysis of tDNA polymorphisms surfaced three additional sequences with the properties of solo LTR sequences, but for which associated CDS were not identified (Table 3.5, S7 Table). Based on these observations and in order to avoid conflicting with the nomenclature proposed for extant full-length elements (Neuveglise, Feldmann et al. 2002) or overlapping with yet-to-be discovered full-length elements for which LTRs have designated numbers, the LTRs discovered in our study were designated LTRyl7, LTRyl8, and LTRyl9. These families grouped by primary sequence and length are characterized by multiple members, and inverted repeat TGT…ACA termini. Most are flanked by short direct repeats of presumed insertion-site sequence.

Non-LTR Ylli LINE retrotransposons

LINE retroelements comprise about 40% of the human genome and exist in many other species (Richardson, Doucet et al. 2015). They are powerful remodelers of eukaryotic genomes because of their ability to transpose other sequences both in trans and cis. LINES were thought absent from Hemiascomycetes until the discovery of the *Y. lipolytica* LINE Ylli in *Y. lipolytica* in 2000 (Casaregola, Neuveglise et al. 2000) and in *Candida albicans* (Goodwin, Ormandy et al. 2001). Ten full-length LINEs were reported in CLIB122 YALI0. In the current study, seventeen full-length elements were identified in CLIB89 YALI1. Much of our understanding of LINES is extrapolated from studies of the active human L1 element (Richardson, Doucet et al. 2015). L1 encodes first and second ORFs that are translated into structural and catalytic proteins, respectively. These associate with cytoplasmic RNAs including genomic Ylli RNA and mediate nuclear re-entry, reverse transcription, and subsequent integration. Ylli similar to L1, is a member of the class of LINES in which ORF2 encodes an apurinic-like endonuclease that nicks chromosomal AT-rich sequences. The DNA 3’ end created by nicking primes reverse transcription of the template genomic RNA starting at the downstream end, a process known as target-primed reverse transcription (Beauregard, Curcio et al. 2008). Reverse transcriptase or repair enzymes complete second-strand synthesis. A distinguishing feature is that reverse transcription is not highly processive so that 5’-truncated LINES tend to accumulate in genomes (Richardson, Doucet et al. 2015). LINE insertions terminate in poly(A) tracts diagnostic of their origin as reverse transcribed RNAP2 transcripts and are typically flanked by short direct repeats of target site DNA.

The *Y. lipolytica* LINE, Ylli, has generic features of LINEs as well as distinguishing characteristics (Table 3.5, S7 Table). Ylli encodes proteins of 714 and 1,300 aa, the second of which contains homology to reverse transcriptase and apurinic-type endonucleases. It exists in multiple 5' truncated copies and is associated with downstream poly(A) tracts (Casaregola, Neuveglise et al. 2002). However, Ylli is distinct from other LINEs in that target-site duplications have not been identified. We speculate that target-site duplications exist, but are too short to be identified, or as sometimes is the case, the ends of the element have non-templated reverse transcribed nucleotides which confound identification of the ends of insertions. Underscoring differences between the CLIB89 and CLIB122 genomes, as mentioned above, there are seventeen apparently full-length, 6.5-kb Ylli insertions in CLIB89 YALI1 and ten in CLIB122 YALI0. In addition, in CLIB89 YALI1, there are seven Ylli sequences greater than 1.0 kb and one hundred and four fragments between 30 bp and 1 kb in length (Table 3.5 and S7 Table).

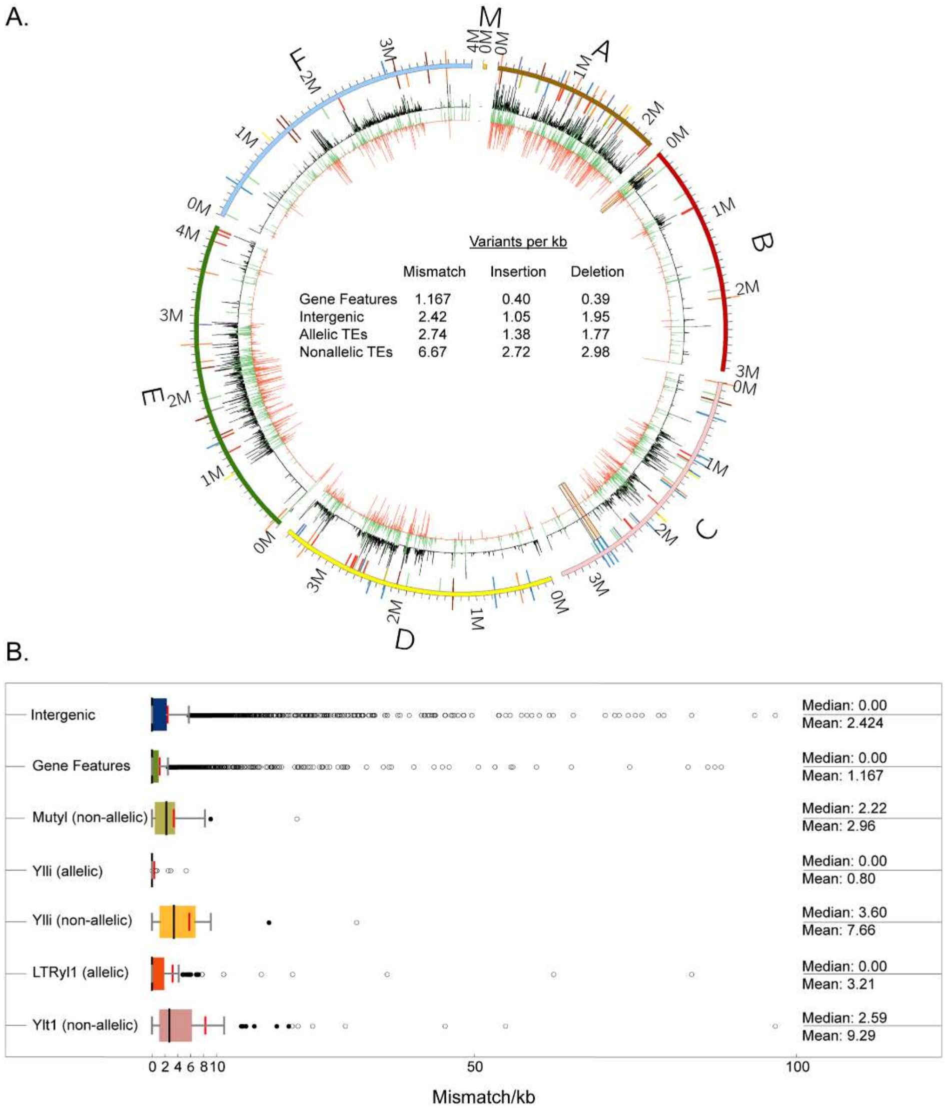
An intriguing feature of metazoan genomes is the expansion of short sequences related to RNAP3 genes referred to as *S*hort *IN*terspersed *E*lements (SINEs) typically about 300 bp in length. Work in human cells has demonstrated that LINEs retrotranspose these RNAs (Dewannieux, Esnault et al. 2003). SINEs are characterized by internal RNAP3 promoter elements in the first segment of the sequence and associated LINE or other RNAP3 sequences in the second half of the element (Vassetzky and Kramerov 2013). Mobilized SINEs are typically associated with downstream poly(A) tracts (Kramerov and Vassetzky 2011). A particularly unique aspect of the *Y. lipolytica* genome is the expansion of tDNAs and dimeric RNAP3 genes relative to other Ascomycetes. This poses an intriguing parallel with the origin of SINEs as dimeric RNAP3 genes. We speculate that retroelement LTRs associated with RNAP3 genes could have provided poly(A) transcript templates for LINE-mediated proliferation of tDNAs in *Y. lipolytica.*

Sources of transposable element differences between CLIB89 and CLIB122

Transposable elements provide for much of the variation within species and are even proposed to account for aspects of speciation. Although CLIB89 was one of the parental strains used to derive CLIB122, we observed striking differences in TE composition. It was of interest to estimate the extent to which segregation versus active transposition contributed to these differences. The abundant Ylt1 in one strain but none in the other strain was particularly striking. We considered three possible hypotheses for the differences between these two closely-related strains in Ylt1 and other elements: 1) there are similar positions and numbers of the element of itnerest in both strains, but ancient versions in the CLIB89 lineage have degenerated and are no longer readily detectable by BLAST analysis; 2) copies were possibly more abundant in CBS6124-2 or CLIB89 and simply segregated differentially; and 3) a wave of retrotransposition sometime after the cross of the *Y. lipolytica* strains resulted in differential proliferation of elements between the two strains.

These hypotheses make distinct predictions regarding the degree of variation in sequence flanking allelic and non-allelic TE. Hypothesis one predicts that elements present in CLIB122 and as relics not identified by BLAST analysis in CLIB89 could be identified by reconstructing the sequence in CLIB122 without the insertion and aligning it to that region in CLIB89. Hypothesis two predicts that elements existing in CLIB122, but not CLIB89, such as Ylt1, would be embedded in sequence inherited from CBS6124-2 and so would be relatively enriched in variants when compared to the same region from CLIB89. Hypothesis three predicts that although some CLIB122 elements might be flanked by variants, if overall an element transposed after the CLIB89 X CBS6124-2 cross, then insertions sites would be randomly distributed across DNA from each parent and *overall* mismatch density would be similar to average genomewide mismatch densities.

To distinguish among these possibilities we took Ylt1 as an example of an element over-represented in CLIB122 and Tyl1 as an example of an element over-represented in CLIB89. We first manually reconstructed 27 full-length element and LTR insertion sites in CLIB122 to derive “naïve” sequences for comparison to CLIB89. BLAST analysis showed that they these empty sites existed in CLIB89, thereby excluding the interpretation that ancient relics in CLIB89 represented insertions identified solely in CLIB122.



Variants per kb

	Mismatch	Insertion	Deletion
Gene Features	1.167	0.40	0.39
Intergenic	2.42	1.05	1.95
Allelic TEs	2.74	1.38	1.77
Nonallelic TEs	6.67	2.72	2.98

Fig 3.5. Genomic differences between CLIB89 YALI0 and CLIB122 YALI1 assemblies (A) Circos diagram illustrating the sequence variation between CLIB122 YALI0 and CLIB89 YALI1 and the locations of annotated TE. Outer circle: TE in both assemblies are represented by colored bars projected outward (CLIB89 YALI1) or inward (CLIB122 YALI0) from the chromosome ideograms: Ylli (blue), Mutyl (purple), Ylt1 (red), LTRyl1 (green), LTRylt7 (orange), LTRyl8 (maroon), LTRyl9 (gold), Tyl3 (yellow), Tyl6 (grey). Inner circles: The black track indicates the mismatch density between corresponding regions between CLIB89 and CLIB122 in each chromosome. The green and orange tracks represent insertions and deletions respectively relative to the CLIB122 YALI0 reference genome assembly. (B) Mismatch density of TE and flanking iLoci of allelic compared to nonallelic elements. The mismatch density (mismatch/kb) of each individual element and flanking iloci is shown. Mean and median (red and black horizontal bars, respectively); near and far outliers (black and white circles, respectively).

In order to more quantitatively evaluate the second and third hypotheses, variant densities (M=mismatch/kb, I=insertions/kb, D=deletions/kb) were quantified and averaged for the four iLoci surrounding the TE insertions (two sequential iLoci per flank) to estimate the extent of differences in the region of the TE. Results of this analysis showed that as expected overall intergenic iLoci averaged higher variant density (M=2.42, I=1.05, D=1.95) than gene feature iLoci (M=1.17, I=0.40, D=0.39)(Panels A and B, Fig 3.5; S8 Table). In contrast, the nonallelic family Ylt1with 10 full length and 17 solo LTR members exhibits flanking regions of even greater variant density (M= 9.29, I=2.25 nd D=3.09) whereas the 30 LTRyl1 allelic elements of the 54 LTRyl1 total elements in CLIB89 YALI1 are embedded in regions of lower variant density (M=3.21, I=1.54, and D=2.52) in contrast to LTRyl1 non-allelic members (M=4.65, I=2.68, and D=1.93) and Ylt1 non-allelic elements. The LINE Ylli showed a similar pattern (non-allelic, M=7.66 versus allelic M=0.80). Furthermore, overall, non-allelic members of TE families occurred in regions of greater variant density (M=6.67, I=2.72, and D=2.98) compared to allelic members (M=2.74, I=1.38, D=1.77)(Panel A, Fig 3.5 Circos; Panel B, Fig 3.5 box plots; S8 Table). The observed differences in the mismatch/kb and insertion densities in comparing all allelic TE to non-allelic TE are significant (p=0.0027 and p=0.01, respectively), but the deletion density difference is less (p=0.082). Overall, therefore while we cannot formally exclude the contribution of TE mobilization subsequent to the CBS6124-2 X CLIB89 cross as an explanation for gross differences between progeny strain CLIB122 and parent CLIB89, data are consistent with hypothesis 2, namely that non-allelic families such as Ylt1 CLIB122 elements were plausibly inherited together with flanking sequence from the CBS6124 parent, rather than supporting hypothesis 3, that subsequent to generation of CLIB122, new insertions occurred and were randomly distributed into intergenic sequence. The analysis developed for this study provides a new tool for differentiating the impact of transposition and chromosomal inheritance to phylogenetic analysis of closely-related strains.

Conclusions

Y. lipolytica is being rapidly adopted as a mainstream species for single cell, biosustainable production of chemicals. In order to provide an improved basis for manipulation of strains related to industrial strain CLIB89, we used hybrid, next-generation sequencing coupled with Irys long-range genome mapping to assemble the CLIB89 genome, making it the first *Y. lipolytica* genome in which single contigs comprise each chromosome. One copy of the rDNA was completely sequenced and tandem repeats of rDNA were documented in the terminal regions of four chromosomes.

Intriguing aspects of the *Y. lipolytica* biology are the greater size of the genome and abundance of RNAP3 species compared to other fungal genomes. We speculate that although identified retrotransposons comprise a relatively small portion of the *Y. lipolytica* genome, they have not only contributed to its expansion but have mediated its adaptation to the diverse environments in which it is found. This possibility is supported by several observations. First, *Y. lipolytica* is striking for the diversity of its elements. These include previously reported DNA element Mutyl, multiple LTR elements and LINE Ylli in full and partial copies. Second, the apparent abundance of LINE fragments and solo LTR-like sequences is consistent with occupation by many now-extinct retrotransposons. Third, autonomous LINES and LTR retrotransposons are present and are known to mediate mobilization of other genomic sequences, including RNAP3 transcripts enabled to evolve into SINEs. We speculate that the surprisingly extensive differences in retrotransposons between related *Y. lipolytica* strains CLIB89 and CLIB122 relate in large part to differences between the parental CBS6124-2 and CLIB89 genomes, and might contribute to inefficient mating between these strains despite mating-compatible mating types.

The current findings further underscore the usefulness of a second, independent *Y. lipolytica* genome sequence. Assembly of a six-contig genome for *Y. lipolytica* will empower ongoing molecular manipulation of this important and novel organism.

Supporting Information

All supplemental files and tables referenced in the text of this chapter can be found at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0162363#sec04

Author Contributions

This chapter was adapted from a manuscript published in PLoS ONE. This work represents a collaborative effort between Christophe Magnan of the lab of Pierre Baldi in the department of Information and Computer Sciences at University of California, Irvine, Ivan Chang, formerly of the Sandmeyer lab, and myself. Specifically, I contributed the genomic DNA that was used for sequencing, the analysis of the

BioNano genome scale mapping, the survey of transposable elements between the different strains, the discovery of the full length Tyl3 element, and PCR confirmation of genome scale structural variants. Christophe Magnan, James Yu, and Ivan Chang are listed as co-first authors in the manuscript. The full manuscript can be read in full here: https://doi.org/10.1371/journal.pone.0162363

CHAPTER 4: Functional genomics for oleaginous yeast *Yarrowia lipolytica*

Introduction

Oleaginous yeast provide a platform for biosustainable production of hydrocarbon-based compounds of industrial and biomedical interest (Beopoulos, Nicaud et al. 2011; Nicaud 2012; Yadav and Stephanopoulos 2014; Qiao, Imam Abidi et al. 2015; Zhu and Jackson 2015). *Yarrowia lipolytica*, an obligate aerobic oleaginous yeast, thrives on diverse carbon sources ranging from plant oils to glycerol and acetate and can be engineered to accumulate up to 90% of its cell weight as lipid (Blazeck, Hill et al. 2014). Additional features of *Y. lipolytica* favoring biotechnology applications include *G*enerally *R*egarded as *S*afe (GRAS) designation (Groenewald, Boekhout et al. 2014), reference genomes (Dujon, Sherman et al. 2004; Magnan, Yu et al. 2016) and expression vectors. Nonetheless, despite metabolomic and transcriptomic studies, tools enabling functional genomics in this system are lacking. The impact of this deficiency is exacerbated by the fact that *Y. lipolytica* is phylogenetically distant from model respiro-fermentative yeasts so that functions of many of the genes cannot be inferred from those of homologs (Dujon, Sherman et al. 2004; Sherman, Durrens et al. 2004).

Functional genomics of the model yeasts *Saccharomyces cerevisiae* (Giaever and Nislow 2014) and *Schizosaccharomyces pombe* is largely based on high-throughput analysis of collections of targeted deletion knockout strains (Winzeler, Shoemaker et al. 1999; Kim, Hayles et al. 2010). However, despite their obvious value, construction of such collections is time and resource intensive. In addition, stringent genetic selection for specific knockout markers can promote enrichment for background suppressor mutations, resulting in a significant frequency of false non-essential classifications (Guo, Park et al. 2013; Teng, Dayhoff-Brannigan et al. 2013). These knockout collections also fail to survey unexpected features, such as intergenic transcripts and short open reading frames (ORFs) (Yoon and Brem 2010; Andrews and Rothnagel 2014). Transposon profiling provides an attractive alternative strategy for production of a mutant collection appropriate for high-throughput and evolutionary analyses (Kumar and Snyder 2001; Crook, Abatemarco et al. 2016).

Approximately one quarter of *Y. lipolytica* genome features have no annotated function while others are assigned functions inferred from homology with genes in model yeasts and other fungi (Dujon, Sherman et al. 2004; Magnan, Yu et al. 2016). Unlike *Saccharomyces* and *Schizosaccharomyces*, *Y. lipolytica* is oleaginous and an obligate aerobe. Even in cases in which *Y. lipolytica* genes have homologs in model yeasts, essential classification can help to differentiate paralogous gene functions. In this study, the Hermes *T*ransposable *E*lement (TE), originally isolated from the housefly *Musca domestica* (Arensburger, Hice et al. 2011), was mobilized for saturation mutagenesis of the *Y. lipolytica* genome. Hermes transposition resulted in over 534,000 independent mutant isolates, or a theoretical average of 1 insertion every 40 bp, based on the genome size of 21 Mbp. We demonstrate the utility of this *Y. lipolytica* transposon library to classify essential and non-essential genes, provide serendipitous insights into gene function, identify non-coding upstream regulatory regions, evaluate genome-scale models, measure contributions of genes under different conditions, and enable reverse genetics by linking phenotypes to specific mutations. To our knowledge, this is the first system enabling *Y. lipolytica* functional genomics.

Results and Discussion

Activation of the Hermes transposon in *Y. lipolytica*

The Hermes transposon system was adapted to *Y. lipolytica* to enable high-throughput transposition and mutagenesis (Materials and Methods). Our results demonstrated efficient mobilization of the Hermes *LEU2* marker in cells that did not incorporate the donor plasmid backbone (Fig S4.1a) and a rare occurrence of double insertions (Fig S4.1b). Analysis of transposon insertion target sites showed that 53% of 535,000 conformed to the canonical insertion site motif 5'nTnnnnAn3' (Guimond, Bideshi et al. 2003; Evertts, Plymire et al. 2007; Gangadharan, Mularoni et al. 2010; Guo, Park et al. 2013; Li, Pettitt et al. 2013) (Fig S4.2a-c) and insertions into intergenic regions were 1.6-fold overrepresented. This observation contrasted with a relatively even distribution of consensus sequences in the genome (FigS4.2d,e). Similar results were observed for Hermes insertions in *S. cerevisiae* and *S. pombe* (Wood, Gwilliam et al. 2002; Alexander, Fang et al. 2010). Analysis of transposon copy number by qPCR indicated that most mutant isolates tested had a single copy of *LEU2* (Fig. S4.1b). Thus, this system has the properties desired for high-throughput mutagenesis of *Y. lipolytica* and linking insertions to phenotypes and gene function. The new Hermes transposon library was utilized to classify essential and non-essential genes, carry out a functional analysis of genes involved in glycolysis and glycerol metabolism, including contributions of genes under competitive growth conditions. We also used our data to compare genome-scale metabolic models for *Y. lipolytica*. In a pilot experiment, we also demonstrated the feasibility of using fluorescence-activated cell sorting to isolate mutant strains with altered lipid metabolism. A transposon profiling workflow diagram is shown in Fig 4.1.

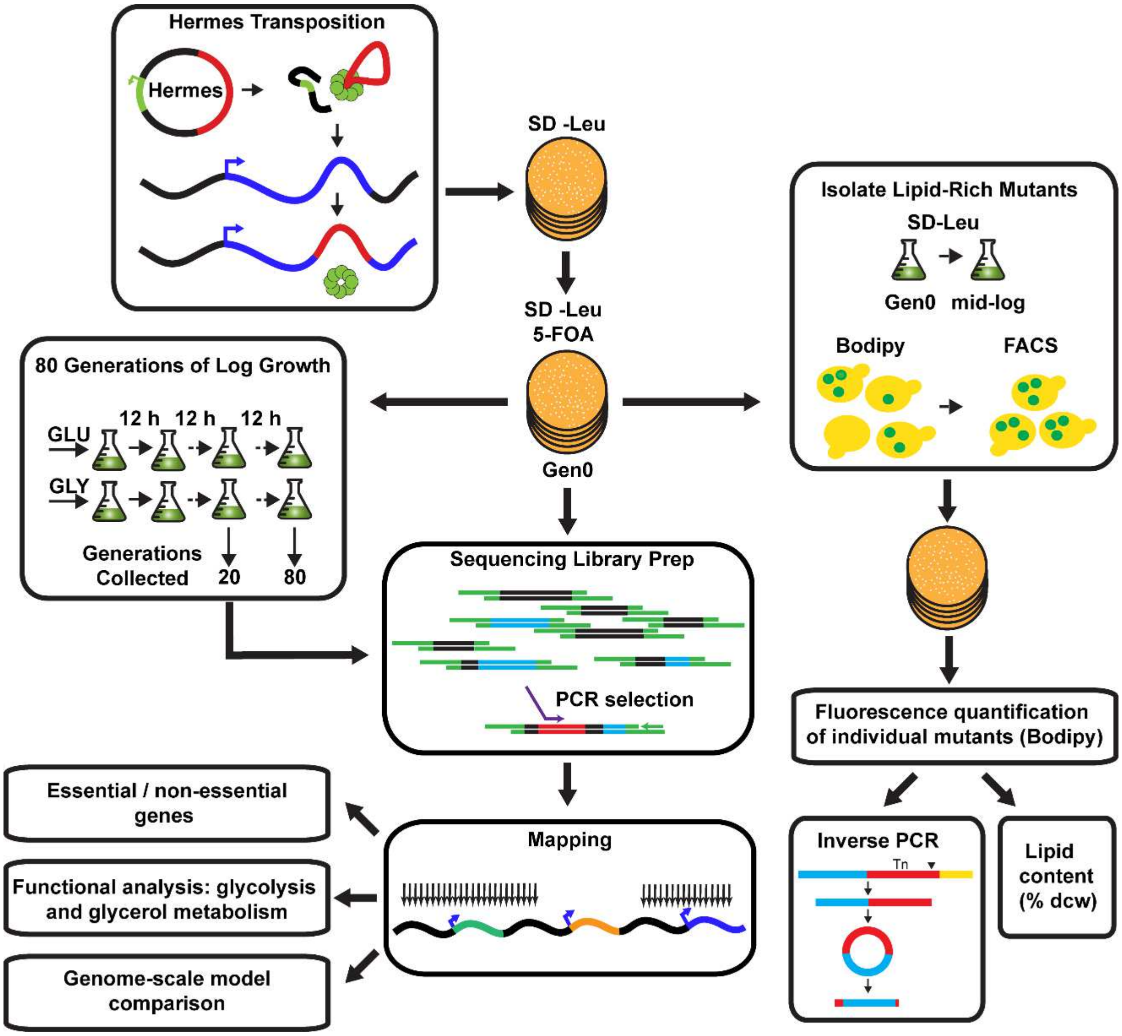

Hermes Transposition
Hermes
SD -Leu
SD -Leu
5-FOA
Gen0
Isolate Lipid-Rich Mutants
SD-Leu
Gen0 mid-log
Bodipy
FACS
80 Generations of Log Growth
12 h 12 h 12 h
GLU
GLY
Generations Collected 20 80
Sequencing Library Prep
PCR selection
Fluorescence quantification of individual mutants (Bodipy)
Essential / non-essential genes
Functional analysis: glycolysis and glycerol metabolism
Genome-scale model comparison
Mapping
Inverse PCR
Tn
Lipid content (% dcw)

Fig. 4.1. Transposon profiling workflow. From top to bottom. Cells were transformed with plasmid (pJY3919) expressing Hermes transposase in *trans* to *LEU2*-tagged Hermes cassette and transformants were collected on SD–Leu medium; colonies were replated to select against donor plasmids and for transposed cells (Gen0). Cells were harvested and regrown ~80 generations in glucose (glu) or glycerol (gly). Cells were sampled at Gen0, Gen20, and Gen80 to identify the distribution of insertions and features were classified according to whether they had expected levels of insertions (non-essential) or significantly fewer (essential). Other studies included functional analysis of genes involved in glycolysis and glycerol metabolism including contributions of genes to competitive growth and comparisons of genome scale models. To isolate lipid-rich mutants, the Bodipy-stained mutant library was sorted by fluorescence intensity as a measure of lipid content into the 1% top and 99% bottom fractions. Individual mutants were recovered and rescreened. Selected mutants were analyzed for lipid content and the location of the Hermes insertion was determined by inverse PCR (Materials and Methods).

Classification of essential and non-essential genes by transposon profiling

In a relatively under-explored yeast such as *Y. lipolytica*, lists of essential/non-essential genes (Table S4.3) streamlines informed genome engineering of host strains, provides insights into metabolic differences with model yeasts, and defines the conditions under which genes are essential to survival. At first consideration, it might seem that cells that have sustained inactivating Hermes insertions into essential genes should not survive to be recovered in the colonies sequenced for insertion sites. However several additional factors influence recovery of insertions. Protein truncation by insertions in the downstream end of an ORF might attenuate rather than eliminate activity, or high nucleosome density could shield genes from insertions and thus reduce insertion density irrespective of essentiality. To score essentiality and incorporate adjustments for these additional inputs, we developed a Poisson-based Gene Classification Strategy (GCS) similar to that described by Gerdes et al. (Gerdes, Scholle et al. 2003). This strategy calculates the probability (P) that Hermes insertions ("hits") within a gene are significantly underrepresented (Materials and methods). Based on instances in which hits occurred exclusively into introns or showed strong bias for the downstream 10% of an ORF, we concluded that such insertions were mostly not inactivating. Such hits were therefore excluded from the final calculation.

Overall, out of 8,710 features (Magnan, Yu et al. 2016), 1962 (22.5%) were classified as essential; 5,907 non-essential (67.8%); and 813 (9.3%) low-confidence essential (LC essential), (Fig. 4.2a, Table S4.3). Hermes insertion profiles over specific genes and flanking sequence (Fig. 4.2b) illustrate patterns typical of essential, non-essential, and LC-essential genes. Essential genes have few to no hits and flanking regions populated with hits. Conversely, non-essential genes have hits in genes as well as flanking regions. Two short LC-essential genes flank *KU70* and illustrate the complications of classifying very short genes. Downstream of *KU70,* YALI1_C11902g predicts a protein with no known homolog or inferred function and sparse flanking hits while upstream of *KU70, MRPL33* (YALI1_C11930g) encodes a homolog of *S. cerevisiae MRPL33,* a member of the large subunit of the mitochondrial ribosome. In *S. cerevisiae*, deletion of this gene prevents respiratory growth (Dimmer, Fritz et al. 2002; Steinmetz, Sinha et al. 2002; Merz and Westermann 2009), so this gene is almost certainly essential in *Y. lipolytica*, an obligate aerobe.

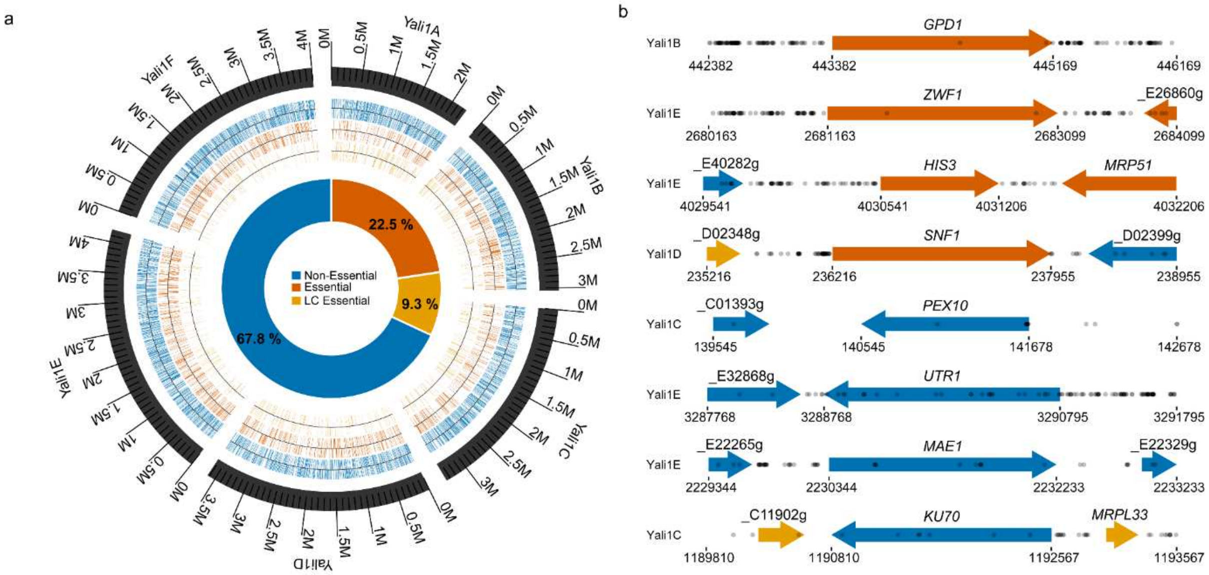


Fig 4.2. Essential and non-essential genes classified by Hermes DNA transposon profiling. (a) Circos diagram showing positions of all non-essential (blue), essential (orange) and low-confidence essential (LC essential; gold) genes. Chromosome size and name are indicated; gene strandedness is indicated by positive (outward) and negative (inward) lines. Central pie chart shows the distribution of classes. (b) Representative loci of essential and non-essential genes. Genes are shown with 1 kb of up- and downstream flanking sequence. Hermes insertion site positions are indicated as grey dots, with regional hit density reflected in the darkness of overlapping dots. Gene names are represented as common names or YALI1 (CLIB89) gene codes positioned above each gene. Arrows indicate strandedness. Essential, genes with few or no insertions; LC essential, genes with relatively few or no insertions but which did not pass the threshold of significance because of size or regional deficit of insertions; non-essential, genes which met the threshold of expected numbers of insertions. Colors are as in (a).

Validation of gene classification method

We evaluated the essential and non-essential classifications of *Y. lipolytica* gene features in several ways. In the first approach, we compared overall results of this classification to those for model yeasts based on systematic knockout collections. The fraction of essential genes (22.5%) in *Y. lipolytica,* was overall comparable to the 20% of ORFs for *S. cerevisiae* (Giaever, Chu et al. 2002) and 26% ORFs for *S. pombe* (Kim, Hayles et al. 2010) called essential.

In a second approach, intragenic hit density or hits/kb/million sample hits (HKM) was plotted for all features. This plot showed a bimodal distribution of intragenic HKM. Poisson classification of essential and non-essential genes separated the distribution into component subpopulations. Essential genes accounted for a significant fraction of very low insertion densities and non-essential genes accounted for a broad continuum of higher insertion densities (Fig. S3a, left panel). In contrast to the bimodal intragenic hit density, a display of intergenic hit density shows a generally unimodal distribution spanning essential and non-essential genes (Fig. S3a, right panel). Thus, genes were not classified as essential as an artifact of anomalous flanking density. Bimodal distribution of hit density for essential and non-essential genes was previously observed for Hermes saturation insertion profiling of the *S. pombe* genome, which was validated by comparison to a knockout collection (Kim, Hayles et al. 2010; Guo, Park et al. 2013). Scatter plots of expected versus observed hits in individual genes show the range of insertions within non-essential and essential gene groups (Fig.S3b).

In a third approach, our classification analysis was tested by engineering disruptions of genes we classified as non-essential and essential. These genes included five non-essential (*GUT2, KU70*, *MAE1*, *PEX10*, and *UTR1*) and three essential (*GPD1, HIS3*, and *SNF1*) genes. As expected, disruption of the non-essential gene set was efficiently accomplished under the conditions used for the initial classification and therefore validated our results. (Table S4.3, Table S4.4). We classified *GPD1* as essential when cells were grown on glucose medium. *GPD1* encodes glycerol phosphate dehydrogenase. In *Y. lipolytica*, it is essential for growth in glucose as a carbon source, but non-essential on medium containing glycerol as a carbon source (Yuzbasheva, Mostova et al. 2017). Cells were disrupted for *GPD1* and were first grown on glycerol as a carbon source. Consistent with previous reports and our essential classification, *GPD1* mutants failed to grow after replica plating onto glucose medium.

Predicted essential genes *HIS3* and *SNF1* were disrupted using CRISPR-Cas9, in the presence of a *URA3*-marked rescue plasmid containing a wild-type copy of the gene harboring a silent mutation that eliminated the CRISPR guide PAM site (Materials and Methods, Fig. S4a). Mutagenized transformants containing the rescue plasmid were colony purified, and disrupting mutations were confirmed by PCR and sequencing as described above. Only wild-type *HIS3 and SNF1* cells were able to grow on medium that selected for cells that lost the respective wild-type rescue plasmid under the same conditions used for the original transposon mutant screen. These results confirmed the essential classification under our experimental conditions (Fig. S4b,c). Further inspection of the data showed that all seven genes for histidine biosynthesis (*HIS1-7*) from phosphoribosyl pyrophosphate to L-histidine are essential. This result was surprising since *HIS3* was previously reported to be disrupted in *Y. lipolytica* (Kretzschmar, Otto et al. 2013).

SNF1 is a master regulator of lipid biosynthesis and disruption mutants are reported to display pleiotropic phenotypes, including elevated lipid production in *Y. lipolytica* (Kretzschmar, Otto et al. 2013; Seip, Jackson et al. 2013). However, we found that mutants encoding a severely truncated gene did not survive in the absence of the *SNF1* rescue plasmid under our growth conditions on glucose medium (Fig. S4c). Others have also noted difficulties in reconciling lipid accumulation and flux analysis properties of different *snf1Δ* mutants (Kerkhoven, Pomraning et al. 2016). A likely explanation of essential classification of *SNF1* and *HIS* biosynthetic pathway genes in comparison to other studies is a significant contribution to viability from unknown genetic factors in our strain background.

Taken together, our results validated the transposon profiling classification for a specific set of genes. In addition, validation enabled unexpected demonstration of nutritional requirements for *Y. lipolytica* strain CLIB89 and showed that *SNF1* and also *HIS1-7* genes of the histidine biosynthetic pathway are essential in this strain background under the growth conditions used in our experiments.

Essential genes in model yeasts are also essential in Y. lipolytica

Essential genes from *S. cerevisiae* (Giaever, Chu et al. 2002; Kastenmayer, Ni et al. 2006; Dowell, Ryan et al. 2010) and *S. pombe* (Kim, Hayles et al. 2010) that have *Y. lipolytica* homologs were assessed for essentiality in *Y. lipolytica*. *S. cerevisiae* and *Y. lipolytica* had 73.4% essential and 9.8% LC-essential genes in common, while *S. pombe* had 69.5% essential and 8.3% LC-essential genes in common (Fig. S5a). To further examine the essential genes by product function, the proportions of essential *Y.*

lipolytica genes that had GO SLIM terms in common with model yeasts was evaluated (Fig. S5b). Of the 1,962 *Y. lipolytica* essential genes, 555 were also essential in both *S. cerevisiae* and *S. pombe*, 137 were in common with *S. cerevisiae* alone and 258 were in common with *S. pombe*. Blast2GO analysis of *Y. lipolytica* essential genes identified several enriched GO SLIM terms. As anticipated, all three yeasts have over-representation of essential genes related to gene expression and replication (Fig S4.5c).

Application: Competitive growth and conditional essentiality

Classification of mutants isolated from colonies provided a relaxed context in which cells grew or died virtually irrespective of doubling time and ability to compete for nutrients in medium containing glucose as the sole carbon source. In order to capture relative contributions of genes to long term batch culture, pooled mutants were passaged during logarithmic growth in rich medium with 2% glucose (YPD) and sampled at Gen0, Gen20 and Gen80. To test for conditional essentiality based on carbon source, the experiment was also carried out in rich medium containing 2% glycerol (YPG). Cultures were sampled and Hermes insertion sites were sequenced at Gen0, Gen20 and Gen80 in order to profile the mutant population over time. Poisson-based GCS analysis was performed (Materials and Methods). Outgrowth of cultures in YPD or YPG reduced diversity of mutant populations (Fig. S4.6, Table S4.7). Despite significant reductions in population diversity for both cultures, the mutant loss was proportional to the distribution of starting mutations (Fig. S4.6b). Interestingly, the majority of mutants that disappeared by Gen80 had intergenic insertions. Essential genes lost proportionately more upstream insertions within 1-kb upstream of the ORF than did non-essential genes in both YPD and YPG. Loss was more similar between these conditions for essential genes (r = 0.84) than for non-essential genes (r = 0.69) (Fig. S4.7), and suggested the loss of mutations in promoters of essential genes was influenced to a greater degree by competitive growth.

The term “growth impaired” was introduced to describe genes that transitioned into the essential class from Gen0 to Gen80 during growth in both YPD and YPG. The terms “Glu- and gly-conditional” describe genes that transitioned in a like manner into the essential class in either YPD or YPG, but not both. By Gen80, 17.5% of genes (1,527) were classified as essential, 8.2% of genes were classified LC essential, and 1.4% of genes were classified as conditionally essential on YPD (0.4%) or YPG (1.0%) (Fig. S6a). The greater number of genes in the gly-conditional class compared to the Glu-conditional class probably reflects the original selection in glucose as the carbon source, but might also reflect additional genes required for growth in glycerol.

Growth-impaired genes (8.2%, 718 genes) were lost in YPD and YPG outgrowth cultures by Gen80. Of the 718 growth-impaired genes, 610 had associated GO terms and of these, 12 were enriched in Gen0 essential while 8 were unique to growth-impaired categories (Fig. S4.6a, Table S4.8). Compared to Gen0 essential GO-enriched genes, Gen80 growth-impaired GO-enriched genes were related to RNA modification, response to heat, peptidyl-amino acid modification and protein acylation.

Application: Contributions of genes to relative fitness

Initial GCS classification of essentiality was based on the absence of mutants representing the respective coding regions. To further assess contributions of individual genes to fitness in competitive liquid culture, we determined representation of mutants in each culture based upon the fraction of reads represented by insertions within each gene relative to all other genes (adapted from (Giaever, Chu et al. 2002; van Opijnen, Bodi et al. 2009; Wiser and Lenski 2015)). We normalized the number of reads per insertion mutation (“hit”) in Gen0, Gen20, and Gen80 glucose and glycerol samples, and derived a gene score (Z-score) reflecting the representation of that mutated gene in each culture (Table S4.3, Materials and Methods). The heatmap display of genes from the Gen0 GCS analysis shows disappearance of mutants in the growth impaired and conditional classes of genes in Glu and Gly cultures, as well as displaying finer distinctions in representation over time (Fig. S4.8). This analysis and other inferences from individual mutations (Table S4.2, Table S4.3, Table S4.9) can be directly applied to strategies for bioengineering of this yeast for growth in liquid cultures.

Application: Comparative and functional analyses of homologs implicated in glycerol metabolism

Relatively few genes were identified as Glu- or gly-conditional. At Gen80, 88 genes were classified as gly-conditional and 34 as Glu-conditional (Table S4.10). Because *Y. lipolytica* is an obligate aerobe, it was expected that genes required for respiration would score as essential or growth impaired, but that there would be some additional genes specifically required for glycerol metabolism. This contrasts with *S. cerevisiae* for which respiration is not essential except for growth on 3-carbon or 2-carbon substrates that do not sustain fermentation. Therefore, it seemed likely that gly-conditional genes in common between *S. cerevisiae* and *Y. lipolytica* would include ones relatively specific for glycerol or gluconeogenesis. Comparison of *Y. lipolytica* Gly-conditional genes to *S. cerevisiae* YPG-essential genes based on growth of the deletion collection in YPG (Qian, Ma et al. 2012) identified nine glycerol-sensitive homologs in common. In addition to genes required for glycerol metabolism, genes related to a variety of other functions were identified as YPG conditional, including 20 genes encoding proteins of unknown function (Table S4.10).

GUT1 (YALI1_F00654g) encoding glycerol kinase, converts glycerol to glycerol-3-phosphate, which is in turn converted to dihydroxyacetone phosphate (DHAP) by mitochondrial glycerol-3-phosphate dehydrogenase encoded by *GUT2* (YALI1_B18499g). DHAP can then be isomerized into glyceraldyde-3-phosphate to enter glycolysis (Fig. 4.3). *GUT1* was essential for growth in YPG (gly-conditional) and *GUT2* was essential for growth in both YPD and YPG medium (growth impaired). The gly-conditional phenotype of *GUT1* and growth-impaired phenotype of *GUT2* is consistent with behavior of independently verified gene disruptions in this work (Table S4.4), and elsewhere in *Y. lipolytica* (Beopoulos, Mrozova et al. 2008) and *S. cerevisae* (Sprague and Cronan 1977). This implies that the Gut1-Gut2 cytoplasmic-mitochondrial pathway operates for growth on glycerol in *Y. lipolytica*.

In a second pathway in many organisms, glycerol can be metabolized in the cytoplasm by oxidation to dihydroxyacetone (DHA) by glycerol dehydrogenase encoded by a member of the broad aldo-keto reductase family of enzymes. In *S. cerevisiae*, this gene is designated *GCY1*. DHA is phosphorylated to DHAP by dihydroxyacetone kinases encoded by *DAK1* and *DAK2*.   The Blastp failed to identify a clear homolog *of GCY1* in *Y. lipolytica*, but identified 12 candidate genes that encode an aldo-keto reductase, annotated in Fig. 4.3 as YPR (Yeast Possible Reductases). One of these YPR genes (YALI1_B20108g) was gly-conditional, which suggests that it may be the aldo-keto reductase primarily associated with the oxidation of glycerol in this alternative pathway. In a similar manner, we identified three proteins with similarity to *DAK1* through BLASTp analysis (Magnan, Yu et al. 2016). However, only YALI1_F02508g was classified as essential, indicating it performs a non-redundant function. Since DHA accumulation is toxic in *S. cerevisiae* (Molin, Norbeck et al. 2003), the cytoplasmic pathway via DAK might also be required despite the alternative Gut1-Gut2 pathway for glycerol metabolism.

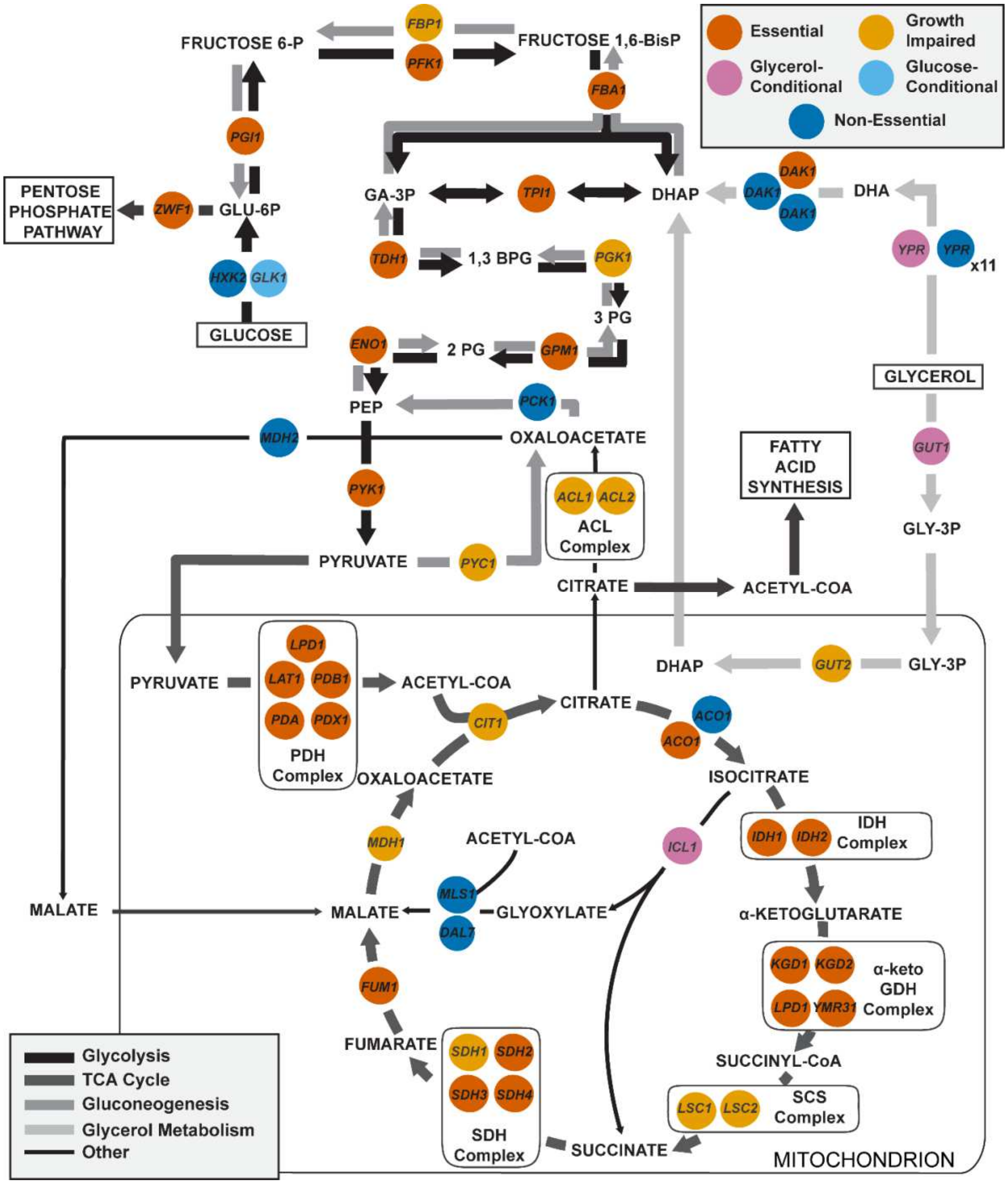


Fig. 4.3. Functional analysis of metabolic genes in glycolysis and glycerol metabolism. Each circle represents a gene color coded by its classification after 80 generations of log phase growth. Genes were classified as essential (orange), non-essential (blue), growth impaired (gold), gly-conditional (reddish purple) or Glu-conditional (sky blue). Abbreviations are as follows: GLY-3p, glycerol-3-phosphate; DHA, dihydroxyacetone; DHAP, dihydroxyacetone phosphate; GLU-6P, glucose-6-phosphate; GA-3P, glyceraldehyde-3-phosphate; 1,3 BPG, 1,3-bisphosphoglycerate; 3PG, 3 phosphoglycerate; 2PG, 2-phosphoglycerate, PEP, phosphoenol pyruvate.

Application: Variable penetrance of TCA cycle mutations

Because *Y. lipolytica* is an obligate aerobe, genes comprising the TCA cycle are expected to be essential. However, it was striking that mutants with insertions in genes that encoded components of the tricarboxylic acid cycle (TCA) cycle survived past Gen0 to be classified as growth impaired (Fig. 4.3, Table S4.11). Overall we identified 21 genes encoding functions associated with the TCA cycle. Of these, 16 are essential and 4 are growth impaired (Fig. 4.3). One possible explanation of the delay in the requirement for *MDH1, CIT1, LSC1, and LSC2* from Gen0 to Gen80 is that the mitochondrial proteins they encode are long lived and sufficient for Gen0 colony development (Guo, Park et al. 2013). This possibility is supported by observations previously made in *S. pombe* for the apparent persistence for many generations of cells with insertion mutations in nuclear genes encoding proteins with essential mitochondrial functions (Guo and Levin 2010). Another possibility is that alternative pathways exist. *LSC1* and *LSC2* encode subunits of the heterodimeric succinate-Co-A ligase complex. This complex is essential in *S. cerevisiae* for conversion of succinyl-CoA into succinate in the TCA cycle (Przybyla-Zawislak, Dennis et al. 1998). However, in animals and certain species of bacteria and yeast, succinyl-CoA-3-oxo-acid CoA transferase (SCOT) can also catalyze this step (Corthesy-Theulaz, Bergonzelli et al. 1997). Instead of removing the CoA group from succinyl-Co-A and generating free CoA, SCOT transfers the CoA group to acetoacetate and generates free succinate and acetoacetyl-CoA. This acetoacetyl-CoA can then be converted back into acetyl-CoA. SCOT is present in *Y. lipolytica* (YALI1_F34029) and was classified as non-essential (Table S4.3). Thus, the non-essential classification of the genes for succinyl-Co-A ligase in Gen0 may be attributable to this alternative route for production of succinate by SCOT.

Application: Identification of strains with altered lipid metabolism

The complexity of the Hermes library enabled screening for desirable phenotypes over a wide range of genotypes. Because of the status of *Y. lipolytica* as an industrial oleaginous yeast, phenotypes related to lipid production and turnover are of interest. Lipid metabolism is influenced by environmental conditions. It has been known for some time that the amount of nitrogen in the growth medium is inversely related to lipid content (Ratledge 2002; Ratledge and Wynn 2002)Ra(Morin, Cescut et al. 2011; Sestric, Munch et al. 2014). In yeasts and filamentous fungi, lipid content is substantially reduced by growth in nitrogen-enriched medium by the control of gene expression and nitrogen catabolite repression (Morin, Cescut et al. 2011; Dulermo, Treton et al. 2013; Seip, Jackson et al. 2013; Wang, Xu et al. 2013; Dulermo, Onodera et al. 2015; Liu, Pan et al. 2015; Kerkhoven, Pomraning et al. 2016; Kerkhoven, Kim et al. 2017; Pomraning, Bredeweg et al. 2017; Wang, Wang et al. 2017; Wang, Li et al. 2018). In *Y. lipolytica*, the mechanisms of nitrogen catabolite repression are not completely understood.

We carried out a preliminary study to investigate the feasibility of screening the transposon mutant library for genes involved in altered lipid metabolism. Screening of the mutant library by flow cytometry indicated that the proportion of lipid-rich cells in the mutant library is ~5-fold greater than that of the control strain when grown in nitrogen-enriched medium, while no fold change was observed under nitrogen-depleted conditions (Fig. S9). To identify mutants with altered lipid regulation in nitrogen-enriched growth conditions, Gen0 cells were cultured in logarithmic growth in nitrogen-enriched medium, stained with Bodipy and sorted by fluorescence-activated cell sorting. Cells were sorted by differential staining into two groups, the 1% most intensely stained cells and the remaining 99% fraction (Fig. 4.1, Materials and Methods). Retest of individual strains from the 1% and 99% fractions by cell cytometry under nitrogen-enriched conditions showed that the 1% fraction had significantly higher Bodipy fluorescence that the 99% fraction. This indicates the effectiveness of the method in isolating potential mutants of interest (Fig. 4.4a, Materials and Methods). Mutants with higher Bodipy staining in nitrogen-enriched conditions may have alterations in genes regulating the flux pathways related to lipid biogenesis, and also have enhanced total lipid content when grown under nitrogen-depleted conditions. To test this hypothesis, several of the highly stained mutants were grown in flask cultures under conditions expected to cause maximal lipid content (nitrogen-depleted medium, 120 hrs at 28ºC) and the total lipid content was determined (Fig. 4.4b, Materials and Methods). Strain yFS2044 had significantly higher mean total lipid content [(26.3% lipid per dcw)] and was 1.7-fold greater than the control (yJY1948). The other strains tested (JY2042, yFS2043, yFS2045, yFS2046) appeared to have higher mean total lipid content, but values were not significant due to variability in the biological replicates. To investigate the nature of the genome disruption in selected strains, the location of Hermes transposon junctions with genomic DNA was identified and mapped to the CLIB89 YALI1 genome (Magnan, Yu et al. 2016) (Fig. 4.4c, Materials and Methods). The transposon insertions in all strains analyzed were located in intergenic regions, consistent with positional bias previously observed (Fig. S3a). In all but one case, the insertion was between two divergently transcribed ORFs such that the expression of one or both genes could be affected. In strain yFS2044, insertion was 380-bp upstream of YALI1_F11261g, encoding a protein of unknown function. BLAST analysis of the protein product of YALI1_F11261g showed the presence of 6 putative WD40 domains. WD40 repeats are highly abundant in eukaryotic genomes and play a role in a variety of cellular processes as either scaffolds for multi-protein complexes or mediators of protein-protein interactions (Schapira, Tyers et al. 2017). In strain yJY2042, the insertion was located 551-bp upstream of *PRB1*, a vacuolar protease. In strains yFS2045 and yFS2046, the total lipid content was not significantly different from control and the effect of insertion between ORFs of unknown function or upstream of 5S rRNA under the experimental conditions is not clear. Nevertheless, the results of this small-scale study demonstrate the potential of expanded investigation using the transposon mutant library in conjunction with Bodipy-staining and cell sorting to identify genes involved in different aspects of lipid metabolism.

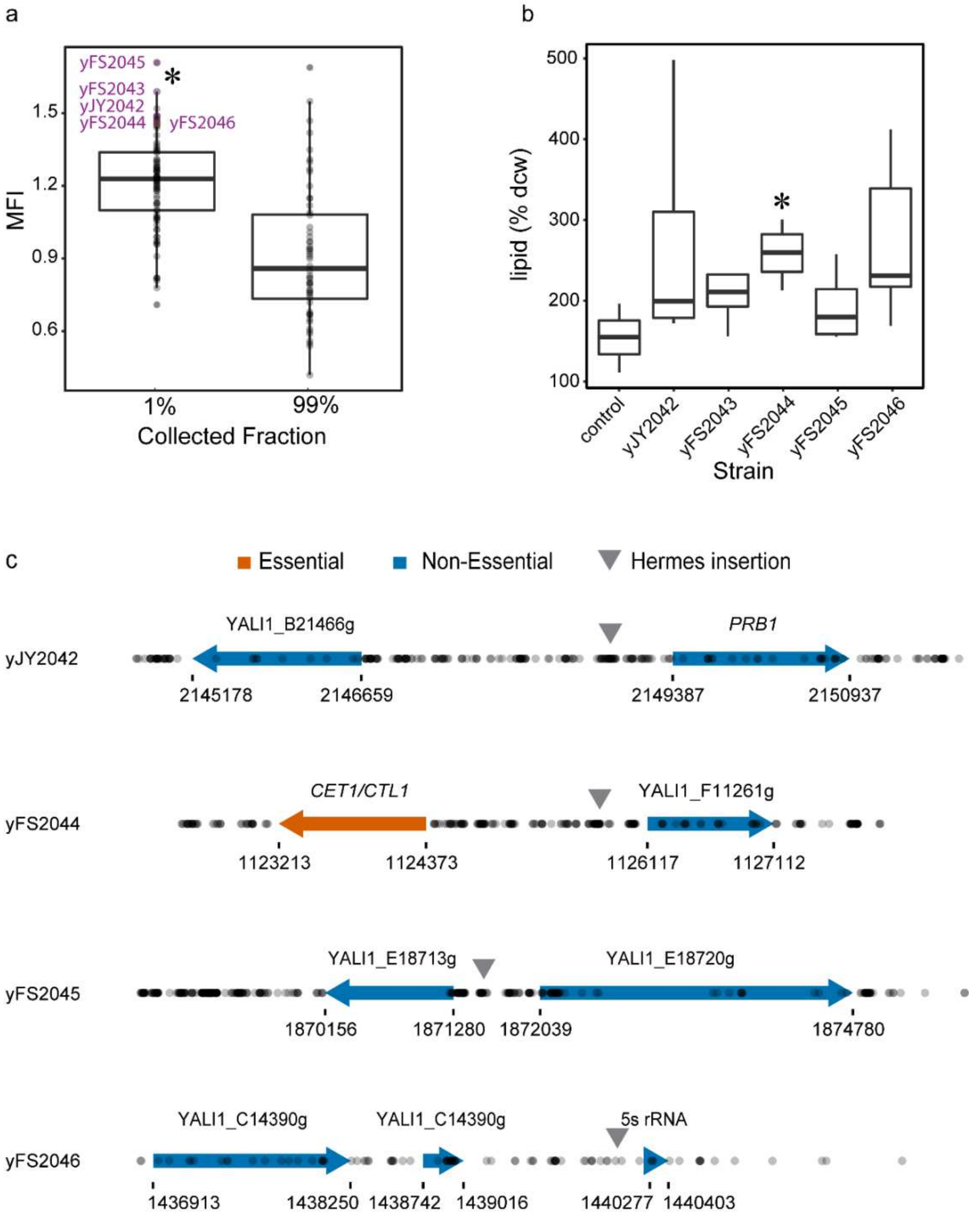


Fig. 4.4 Isolation of lipid-rich mutants from Bodipy-stained mutant library using cell sorting. (a) Retesting of individual Bodipy-stained mutant cells from the original FACS-collected fractions (top 1% and remaining 99%) using NovoCyte cytometer. Boxplot of relative lipid content of individual Bodipy-stained mutants measured as normalized geometric mean fluorescence (MFI) in comparison to the control strain. Cells were grown in nitrogen-replete medium. Dots represent individual strains tested. Mutant strains that were subjected to further analysis are labeled. *Indicates significant difference in MFI. (b) Lipid content [% lipid per dcw] of selected mutants grown in nitrogen-depleted medium for 120 hrs. Replicates n=4 or 5. *Indicates significant difference. (c) Position of Hermes insertion in selected mutant strains.

Application: Functional genomics with in silico network modeling

Transposon profiling-based functional genomics is most useful if integrated with other genome-wide information for understanding biological processes. Genome-scale metabolic models (GEMs) are reconstructions of metabolic pathways at the systems level (Becker, Feist et al. 2007). Such models serve as frameworks for integrating multiple levels of heterogeneous data and so provide mechanistic insights into the relationship between genotype and phenotype. GEMs are constructed by linking genome annotations to reaction databases such as KEGG, thereby establishing a reaction network corresponding to known and hypothetical genes of an organism (Kanehisa and Goto 2000). GEMs can be combined with constraint-based linear/non-linear programming algorithms to predict the biomass growth rate of the organism given a particular growth condition. The standard way to measure the accuracy of and train GEMs is to compile a library of gene knockout phenotypes from experimental results, generate a “truth table” of true positives, false positives, true negatives, and false negatives, and compare with the GEMs *in silico* predicted phenotypes. Hermes transposon profiling provided whole-genome information of knockout phenotypes such that all modeled genes and the reactions to which they map can be tested to evaluate and train the models (Fig. 4.5a, Table S4.12). We chose two GEMs for *Y. lipolytica*, iNL895 (Loira, Dulermo et al. 2012) and iYALI4 (Kerkhoven, Pomraning et al. 2016) to test performance against the classification of essential and non-essential genes determined by GCS. We sought to compare the overall performance of the two models. In addition, we tested for systematic weaknesses in the models by identifying GO groups that were enriched for genes where the models failed to correctly distinguish inviable and viable phenotypes. The iYALI4 and iNL895 models converged at about 80% true positives and 40% true negatives, meaning that they underestimated the fraction of essential genes. Specific comparison of the models to the transposon classification showed that the iNL895 model and the iYALI4 model similarly predicted a little less than half of the functional profiling prediction, with slight differences related to a small number of genes that differed between the two models (Fig. 4.5b). The Receiver Operator Curve (ROC) showed no significant difference between the models (Fig. 4.5a). The models are in part based on *S. cerevisiae* GEM (Duarte, Herrgard et al. 2004; Nookaew, Jewett et al. 2008). Examination of GO terms that were enriched for genes that were for the most part under-identified showed that mitochondrial genes including tRNA synthetases were among those processes (Table S4.12).

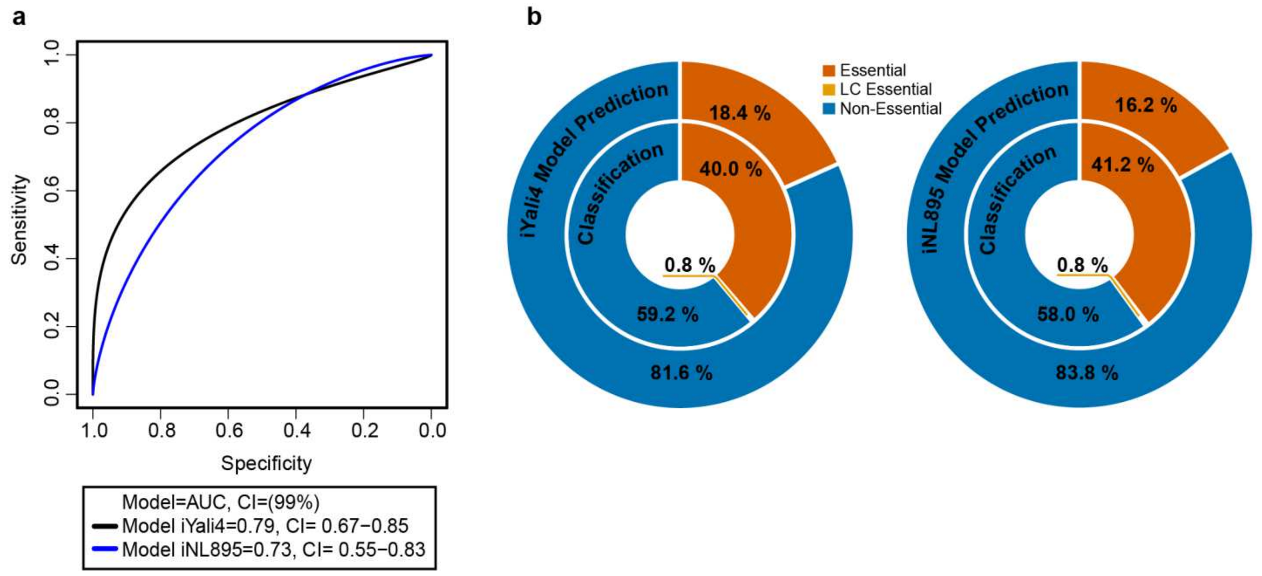


Fig. 4.5. Comparison of genome scale metabolic models for *Y. lipolytica*. The iNL895 (Loira, Dulermo et al. 2012) and iYALI4 models were evaluated using essential, non-essential and LC-essential classifications based on Hermes transposon profiling. (a) Models were allowed to execute overall genes at a thresholds of 0.5 and predict resulting biomass of single gene deletions. Receiver Operator Curve (ROC) shows performance of model biomass predictions using the Hermes transposon profiling essential/non-essential classifications as binary classifiers. ROC curves were evaluated using the area under the curve (AUC) and 99% confidence intervals (CI). Both curves show better than random (slope = 1) performance, however the two curves are not significantly different from each other (bootstrap p-value = 0.4756). (b) Comparison of percentage of modeled genes classified by Hermes insertion profiling (inner circular) to performance of same genes classified by either iYali4 model (left) or iNL895 model (right). Gene color: essential (orange), LC essential (gold), and non-essential (blue).

Conclusions

This work enabled functional genomics in *Y. lipolytica* and underscored the advantages of a system of random insertion mutagenesis. Strikingly, it demonstrated the extent to which gene contributions to fitness fall into a continuum: some are absolutely essential for survival even within pure mutant colonies under the most permissive conditions, but a significant proportion of “non-essential” genes are required for cells to compete successfully in mixed cultures or under specific conditions. Finally, some insertions led to truncated essential proteins and viable phenotypes, suggesting that hypomorphic mutations can be rationally generated to study essential functions.

Over a half-million Hermes insertion mutants were derived, allowing a high degree of granularity for assessing *Y. lipolytica* gene contributions to fitness. Essential genes were defined as those that had significantly less than the expected number of insertions after normalizing for gene length and flanking insertion frequency and filtering to remove intronic or 3’-bias. GCS was validated by comparisons to model yeasts, examination of conserved gene families, expectations based on known respiratory physiology of *Y. lipolytica*, and directed knockouts. Validation by these independent criteria showed that the method was sufficiently robust to draw useful inferences regarding the biology of this interesting organism.

The essential gene census derived here has multiple applications. First**,** on a practical level, random insertion patterns indicated that 22 to 30% of genes make critical contributions to survival and fitness. Attempts to attenuate function of these critical genes will now be better informed. Second, identification of unexpected non-essential genes highlighted redundancy of enzymes in the TCA cycle, and essentiality of glycerol catabolic pathways. Third, the list of essential genes will potentiate future systems biology of *Y. lipolytica* as illustrated through testing of GEMs and identification of classes of functions where the model does not perform well. Fourth, we highlighted the challenge for future investigations of the 5% of genes with no known homolog or even recognizable domain that are nonetheless essential.

We also highlight applications of quantifying the contributions of non-essential genes to growth. The calculation developed for relative mutant fitness differed from that described for essential gene determination in that it reflected representation of mutants in the cycling populations. This analysis identified mutants that were recovered in the Gen0 insertion library but were challenged in the mixed culture growth regime. An interesting case was presented by mitochondrial genes that are almost certainly essential for survival, but may have naturally long half-lives. The charts of fitness can also serve as a guide to whether knockouts or specific mutants require careful interpretation, since in some instances the diversity of mutations within a single gene collapsed during growth from tens of insertions down to only a few. In contrast to deletion collections in which non-essential genes are represented by single mutants, the starting diversity of Hermes insertion mutations within genes can serve as a more complete guide for bioengineering to produce attenuated as well as knock-out phenotypes of both essential and non-essential genes.

One of the benefits of random insertion mutagenesis is recovery of phenotypes from insertions in noncoding regions, dubious ORFs or other unexpected features that can be linked to specific insertions. In a small-scale study, we demonstrated the feasibility of using fluorescence-activated cell sorting to identify mutants in lipid metabolism, thereby underscoring the potential of transposon profiling libraries to identify regulatory mutations.

The genomics resource described here will complement the expanding toolbox for development of *Y. lipolytica* as an industrial microbe. This Hermes transposon is portable to other *Y. lipolytica* strains and can be activated in other genetic backgrounds to identify synthetic phenotypes, dissect other biological processes, or, when coupled with fitness analysis, identify strains with robust survival. Genomewide profiling in *Y. lipolytica* will complement ongoing metabolomics, flux analysis and modeling studies to further enable the development of this unique single cell respiratory model system.

Author Contributions

This chapter is adapted from a manuscript that is currently in revision/review. This work represents a collaborative effort between me and Kurt Patterson, formerly of the Sandmeyer lab. My contributions to this work involves the majority of the molecular work, including the construction of the strains and plasmids used in this study, development of an improved transformation protocol, validation of the essential gene classifications using CRISPR, analysis of the metabolic pathways, especially of the TCA cycle, and the planning and analysis of the follow up experiments. Kurt Patterson and James Yu are listed as co-first authors in this manuscript.

CHAPTER 5: A promising pipeline from polyketide to pogostone in *Yarrowia lipolytica*

The nonconventional yeast *Yarrowia lipolytica* has emerged as a model host organism for the biosustainable production of industrial chemicals from renewable carbon sources. It is most notable for its ability to accumulate high levels of lipids under nutrient limiting conditions, but recent studies have shown that it is also a suitable host for efficient production of other industrially relevant metabolites, such as free fatty acids, organic acids, and polyols (Beopoulos, Cescut et al. 2009; Tomaszewska, Rywinska et al. 2012; Liu, Chi et al. 2013; Xu, Qiao et al. 2017; Yuzbasheva, Mostova et al. 2017; Zeng, Zhang et al. 2017).

Another class of natural metabolites with industrial and biomedical applications are polyketides. Triacetic Acid Lactone (TAL, or 4-hydroxy-6-methyl-2-pyrone) is a polyketide that is synthesized by the 2-pyrone synthase (2-PS) gene, originally isolated from *Gerbera hybrida* (Austin and Noel 2003). TAL synthesis via 2-PS is similar to fatty acid synthesis: an acetyl-CoA starter molecule and two malonyl-CoA molecules are iteratively condensed. TAL has many applications as a platform molecule that can be differentiated into a range of commodity chemicals for use as food additives (sorbic and hexanoic acid), antibiotics (pogostone), and fuel additives (acetylacetone) (Chia, Haider et al. 2013; Kraus, Wanninayake et al. 2016).

Microbial synthesis of TAL has been conducted in both *Escherichia coli* and *Saccharomcyes cerevisiae* (Tang, Feng et al. 2013; Cardenas and Da Silva 2014; Saunders, Bowman et al. 2015). Previous studies demonstrated that TAL is toxic to *E.*

coli (via personal communication with Laura Jarboe), thus limiting its feasibility as a host microbe for the production of TAL. A more recent study expressing the 2-PS enzyme in industrial strains of *S. cerevisiae* showed titers up to 5.1 g/L (Saunders, Bowman et al. 2015). However, this work was conducted in complex YPD medium, which has bioactive constituents that inhibit the downstream processes needed to convert this metabolite into sorbic acid (Schwartz, Johnson et al. 2014).

Our study represents the first to use *Y. lipolytica* as a host organism for the production of polyketides. Since both fatty acid synthesis and TAL synthesis utilize identical acetyl-CoA and malonyl-CoA substrates, *Y. lipolytica* may be an ideal host for the production of polyketides like TAL. In this study, we demonstrate that conditions favoring lipid accumulation also favor TAL accumulation in *Y. lipolytica*. In addition, the TAL produced in this growth media composition can then be crystallized in a form with high purity that is suitable for the downstream production of pogostone.

The ATP Citrate Lyase (ACL) pathway, which is unique to oleaginous organisms like *Y. lipolytica* is suggested to be the primary source of *de novo* acetyl-CoA for the production of fatty acids in nitrogen limitation (Boulton 1981; Blazeck, Hill et al. 2014). This pathway has been implicated to provide the cytosolic acetyl-CoA necessary for *de novo* fatty acid synthesis and eventual lipid storage. Previous studies have shown that in nitrogen-limited (NL) conditions, wild type strains of *Y. lipolytica* are capable of accumulating up to 36% lipid and up to 81.4% lipid with additional genetic modifications (Beopoulos, Cescut et al. 2009; Xu, Qiao et al. 2017).

Since acetyl-CoA is required for TAL and fatty acid synthesis, we asked whether nitrogen limitation may also lead to increased TAL accumulation. An auxotrophic strain of W29 *Y. lipolytica* (yJY2006), was transformed with a plasmid containing 3 copies of the 2-PS gene driven by the *Y. lipolytica TEF1* intron promoter (pDTN3997) and was cultured in both NL medium and nitrogen-replete medium (NR) that is NL medium supplemented with 5 g/L ammonium sulfate, a source of nitrogen for *Y. lipolytica*. Cultures in NL medium accumulated 1.4 fold more TAL than NR cultures (Fig. 5.1a), suggesting that acetyl-CoA generated from the ACL pathway contributes acetyl-CoA to heterologous TAL synthesis. This mechanism also suggests that restricting fatty acid synthesis would redirect acetyl-CoA into TAL synthesis. To investigate whether the acetyl-CoA synthesized from the ACL pathway under nitrogen limitation can be further diverted from fatty acid synthesis into the TAL pathway, cerulenin, a potent inhibitor of fungal fatty acid synthase, was added to the medium (Omura 1976). yJY2006 was cultured in flasks containing NL medium. After 27 hrs of growth, the cultures were split into two sets, and a final of concentration of 20 μM cerulenin was added to one set and an equivalent volume of DMSO to the other. After an additional 73 hrs of growth, cerulenin-treated cultures accumulated 2.7 fold more TAL than in NL medium (Fig. 5.1b). Taken together, these two findings are consistent with the model that acetyl-CoA generated in nitrogen-limiting conditions can be diverted away from fatty acid synthesis and into TAL synthesis.

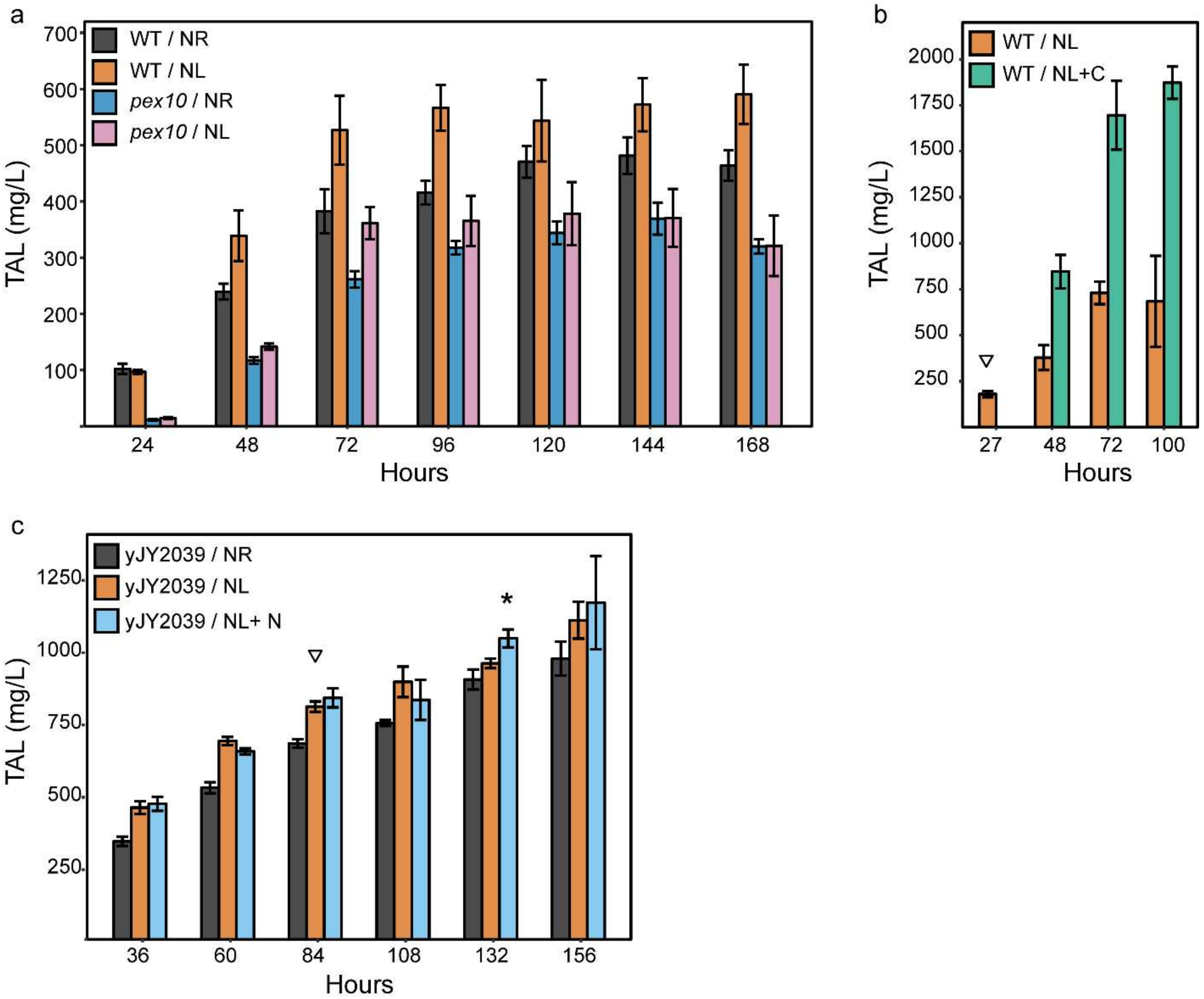


Fig. 5.1. TAL production in *Y. lipolytica*. (a)TAL concentration in flask-grown cultures in nitrogen enriched (NR) and nitrogen limited (NL) medium. Comparison of WT (yJY2006) and *pex10*. (b) Effect of cerulenin on TAL production in WT (yJY2006) cultures in NL medium. Cerulenin added at 27 hrs (triangle). (c) Nitrogen addition increases TAL production in NL culture. TAL concentration in NR, NL or NL medium with additional nitrogen added (NL + N). Nitrogen added at 84 hrs (triangle). * indicates significant difference.

We have shown that the acetyl-CoA produced under nitrogen limitation can be used for TAL synthesis. However, acetyl-CoA is also generated by beta oxidation of fatty acids. In *S. cerevisiae*, beta oxidation occurs exclusively in the peroxisomes (Kunau, Dommes et al. 1995). To investigate the contribution of peroxisomal beta oxidation to acetyl-CoA as a substrate for TAL synthesis, we utilized the CRISPR/Cas9 system to disrupt *PEX10* (YALI0C01023g/ YALI1C01416g) in *Y. lipolytica* to make the strain yJL2008 (Patterson et al., 2017). Pex10p is an integral membrane protein essential for the formation of peroxisomes, thus yJL2008 will not form functional peroxisomes to carry out peroxisomal beta oxidation. When transformed with the 2-PS expression plasmid (pDTN3997), disruption of *PEX10* decreased TAL titers by 66% over the control strain in NL conditions, suggesting that a basal level of beta oxidation exists within cells even in conditions conducive for lipid accumulation (Fig. 5.1a).

This result is also consistent with the model that acetyl-CoA produced from beta oxidation can be used to synthesize TAL. Since acetyl-CoA generated from beta oxidation was shown to be utilized for TAL synthesis, we tested whether cells that have accumulated lipid in NL conditions can generate additional TAL by re-stimulating cell growth and promoting the oxidation of stored lipid bodies. An additional 4.35 g/L ammonium sulfate was added to NL medium after 84 hrs of growth to stimulate cell growth after the lipid accumulating stage. An additional 48 hrs of culturing increased TAL titers by an additional 1.2 fold (Fig. 5.1c).

In contrast with *S. cerevisiae*, in the mammalian system beta oxidation occurs in both mitochondria and the peroxisomes. However, recent studies have demonstrated that other fungi are also capable of mitochondrial beta oxidation, opening the possibility that *Y. lipolytica* and other fungal species that have diverged from *S. cerevisiae* may also have functional mitochondrial beta oxidation (Maggio-Hall and Keller 2004; Hynes, Murray et al. 2008; Kretschmer, Wang et al. 2012).

A subsequent BLAST search of the *Y. lipolytica* genome identified a putative acyl-CoA dehydrogenase (YALI0D15708g/YALI1D19252g) and a putative enoyl-CoA hydratase (YALI0B10406g/YALI1B14044g). Both of these enzymes are required for mitochondrial beta oxidation, providing further suggestive evidence of the existence of this pathway in this organism. Both the finding that there is a constitutive level of peroxisomal beta oxidation and the existence of a mitochondrial beta oxidation warrant further investigation in *Y. lipolytica*. Indeed, beta oxidation may serve as a source of acetyl-CoA that may have significant industrial implications in *Y. lipolytica*.

To further improve titer and yield, we cultured the TAL-producing strain in a bioreactor in batch fermentations. yJY2039 is a strain derived from yJY2006 with the 2-PS expression cassette integrated into the genome. When transformed with 2-PS expression plasmid (pDTN3997) to give additional copies of 2-PS, yJY2039 produces 1.4 fold more TAL than yJY2006 in NL medium.

In order to replicate the nitrogen limitation and subsequent nitrogen replenishment in fermentation conditions that produced high levels of TAL in the flask, yJY2039 was first inoculated into a bioreactor containing NL medium supplemented with 40 g/L glucose. After 40 hrs of growth, 4.35 g/L ammonium sulfate was added to the medium to induce beta oxidation, and the base was switched from sodium hydroxide to ammonium hydroxide. The strain was cultured for an additional 40 hrs before the medium was collected and analyzed for TAL production. These batch fermentations yielded titers of 2.6 g/L TAL, which is 12% maximum theoretical yield.

TAL can serve as a precursor of pogostone and its analogs that are important molecules with antifungal and antibacterial activity (Swamy and Sinniah 2015). Previous syntheses of pogostone reacted dehydroacetic acid with isobutryaldehyde in an aldol reaction followed by hydrogenation to give pogostone. The best yield obtained for pogostone was 58%. (Tang, et al. 2015).

After studying the pogostone structure we devised a synthetic strategy to derive pogostone from TAL in a one-pot procedure (Fig. 5.2a). Our synthesis began with commercially available TAL and 4-methylpentanoic acid. *N,N′*-Dicyclohexylcarbodiimide (DCC) and 4-dimethylaminopyridine (DMAP) are mild reagents for the preparation of esters and amides (Neises and Steglich 1978). Using these two reagents we achieved a one-pot preparation of 3-acyl-4-hydroxy-2-pyrones from carboxylic acids and TAL. This reaction proceeds through a Fries type rearrangement of the O-enol acyl group to the α position of the lactone to get the desired C-acylation product (Fig. 5.2a). A similar yield of pogostone was also obtained using TAL purified from two biocatalysis-derived samples which differed by their purification methods from the biomass broth as 100% and 94% purity (Supplemental Information). When we subjected these samples for the one-pot synthesis protocol, we were able to get between 93-99% yield for the 100% pure TAL and the 94% yield from the 94% purity TAL. A modification of the same reaction protocol (Fig. 5.2b) was used to produce six different analogs of pogostone, including isobutyric acid, isovaleric acid, 6-methylheptanoic acid, 3-methypentanoic acid, heptanoic acid and 5-methylhexanoic acid. The isolated yields were 58%, 87%, 85%, 99%, 87% and 88% respectively. (Table S3, Supplemental Information).

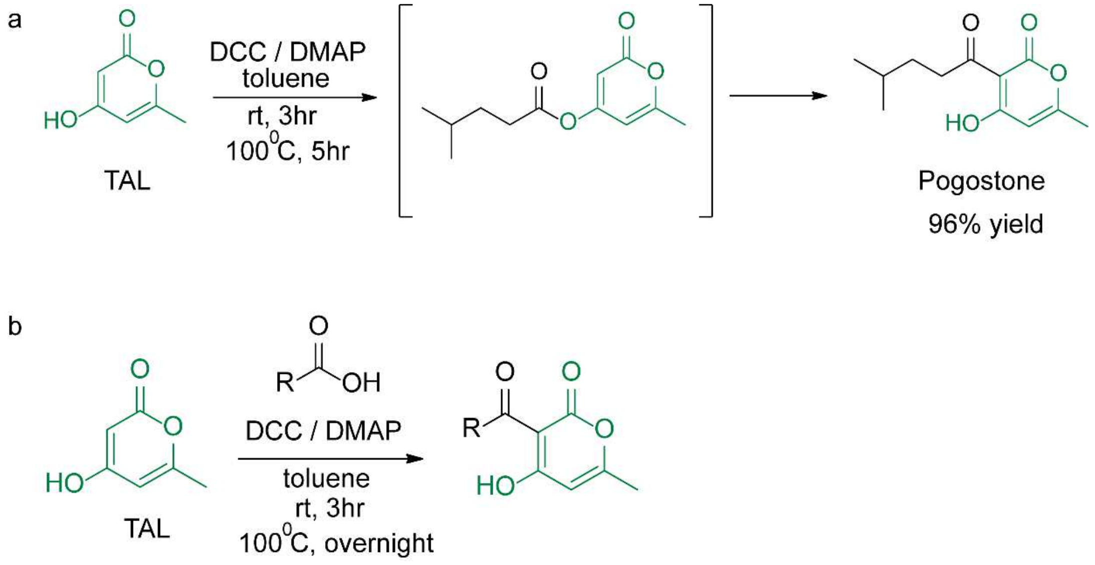


Fig. 5.2. One-pot protocol to synthesize pogostone and its analogs from TAL. (a) Scheme for synthesis of pogostone. (b) Scheme for synthesis of pogostone analogs from TAL.

Finally, an unexplored advantage to utilizing *Yarrowia lipolytica* as a host microbe is its ability to grow robustly in minimal medium. An underappreciated advantage of utilizing an organism that is typically cultured in synthetic medium is the relative simplicity of isolating TAL from the medium. Previous efforts in industrial metabolic engineering have been focused on yields and titers with little to no regard to the purification processes and catalytic reactions necessary to convert a naturally produced metabolite into a commodity chemical of interest. In contrast, we demonstrate that the renewably sourced TAL produced in our minimal medium formulation can be isolated with high purity and yields and with little to no bioactive contaminants that would interfere with the reactions necessary to convert TAL into pogostone. The crystallized TAL isolated from the medium is in a form that is conducive for all downstream processes, not just conversion to pogostone.

Yarrowia lipolytica is widely recognized for its ability to accumulate high levels of lipids, indicating that it has high carbon flux into acetyl-CoA. Our work demonstrates that this high flux into acetyl-CoA may be diverted away from fatty acid synthesis and that this organism, and other oleaginous organisms with the ACL pathway, may serve as ideal hosts for the production of polyketides and other acetyl-CoA derived metabolites.

CHAPTER 6: Summary and Conclusions

Developing new technologies to accelerate the bioengineering of *Y. lipolytica*

At the onset of this work, one of the most striking deficiencies missing from the molecular toolkit of *Y. lipolytica* was the lack of a complete genome assembly. To address this, we proposed to build an independent genome assembly of *Y. lipolytica* using traditional short read Illumina technology coupled with contemporary PacBio long read and BioNano genome mapping technologies. This combinatorial approach not only allowed us to assemble the genome into chromosome length contigs but also allowed us to accurately place repetitive sequences and identify intriguing genomic features such as a 54 kb block repeat and several overlapping reading frames.

Subsequent comparisons of the CLIB89 genome with the previous CLIB122 genome assembly identified several striking differences between the genomes of these two closely related strains of *Y. lipolytica*. The Ylt1 and Tyl6 families of Ty3/Gypsy elements are completely absent from CLIB89. Genomic locations of full length Ylli retroelements differ between the two strains, suggesting that this element may still be actively transposing. These findings highlight the importance of accurate genome assemblies for industrial strains of *Y. lipolytica*; assumptions and observations made from the genome assembly of one strain may not be applicable to the other strains. For instance, integration strategies that capitalized on zeta repetitive elements (Ylt1 solo LTRs) would not function in strains lacking the Ylt1 retroelement (Schmid-Berger, Schmid et al. 1994; Fickers, Fudalej et al. 2005).

After assembly of a high quality genome, we sought to identify the essential genes in this organism. To accomplish this goal, we performed saturation mutagenesis with the Hermes DNA cut-and-paste transposable element. Sequencing of all the Hermes insertion sites revealed notable gaps in the genome which we surmised to be essential genes. By mapping these genes onto metabolic pathways, we were able to also assess the necessity of entire pathways. For instance, we discovered that the entire histidine biosynthetic pathway was essential. This pathway also includes *HIS3*, one of the most commonly used nutritional markers in *S. cerevisiae*. Findings such as these may help gain insight into the nutritional requirements necessary for growth in *Y. lipolytica* which may have consequences for future bioengineering of this organism.

A list of essential genes would not only inform and direct future metabolic engineering efforts but would could also be used to assess the validity of genome scale metabolic models. Several studies advocate for the use of *in silico* predictive models, especially for metabolic enzymes that may have a multi-tiered system of transcriptional, translational, and post translational control of expression (Loira, Dulermo et al. 2012; Kerkhoven, Pomraning et al. 2016; Kerkhoven, Kim et al. 2017). Using our essential gene list, we demonstrate that the currently existing metabolic models underestimate the number of essential genes to a similar extent. The development of future metabolic models would be made more accurate by incorporating the essential gene data set identified by our transposon mediated functional genomics approach.

In the course of our work, we also developed several protocols to stably integrate heterologous expression cassettes into the *Y. lipolytica* genome. Unlike *S. cerevisiae* which possesses a very high rate of homologous recombination, *Y. lipolytica* has a demonstrably low rate of targeted integration, even with very long 1 kb flanking sequence. Although exogenous DNA is efficiently incorporated into the genome, integration via non homologous end joining results in random integration sites, which may or may not be favorable for high levels of transcriptional expression. In response, we developed a transposition based integration strategy utilizing the Hermes DNA cut-and-paste transposon. The Hermes transposon favors integration into regions with low nucleosome occupancy, which are regions typically associated with active transcription and represents a promising alternative to random integration. Another alternative integration method would be to use CRISPR to target repetitive elements in the genome. We demonstrate the feasibility of this strategy by utilizing a CRISPR construct that targeted the repetitive Ylli retroelement. This strategy has the potential to quickly insert multiple copies of a gene of interest, multimeric enzyme complexes, or even entire multi enzyme pathways into the *Y. lipolytica* genome. However, in order to assess the efficacy of these alternative integration strategies, GFP can be placed under the control of a *Y. lipolytica* promoter and integrated into the genome by one of these methods. Fluorescence intensity could be quantified with cytometry and, coupled with real time PCR to quantify copy number, the level of expression per copy number for each of these integration strategies could be compared.

Bioengineering of *Y. lipolytica* for the production of a novel polyketide

Many groups have utilized Y. lipolytica and other oleaginous organisms for the microbial production of lipids in nitrogen limited medium. It has been speculated that the ATP citrate lyase pathway is activated upon nitrogen depletion, and that the resulting cytosolic acetyl CoA is directed towards *de novo* fatty acid synthesis and lipid production. We have shown that under nitrogen limitation, *Y. lipolytica* also shows increased TAL production. This result suggests that the acetyl CoA produced in this pathway may be redirected into other pathways that also utilize acetyl CoA. Additionally, the requirement for nitrogen limiting conditions precludes the use of complex medium, which has been shown to be detrimental to the downstream conversion to sorbic acid (Schwartz, Johnson et al. 2014). We show that TAL can be purified from the minimal growth medium in a form that is suitable for downstream conversion into pogostone.

Recent studies have reported that there is particularly strong lipid turnover in wildtype *Y. lipolytica*, and stored triglycerides are oxidized and citric acid and mannitol are produced, even in the complete absence of nitrogen (Papanikolaou, Chatzifragkou et al. 2009; Makri, Fakas et al. 2010; Tchakouteu, Kalantzi et al. 2015). These reports are consistent with our finding that acetyl CoA resulting from constitutively active peroxisomal beta oxidation significantly contributes to TAL titers. In response, several groups have engineered strains of *Y. lipolytica* that are incapable of beta oxidation and these engineered strains are capable of accumulating up to 90% lipid (Blazeck, Hill et al. 2014). Taken together, these results demonstrate the potential flexibility of a *Y. lipolytica* microbial factory. With proper engineering, *Y. lipolytica* can be an efficient host organism not only for the production of microbial lipids, but also for polyketides and other acetyl CoA based molecules.

The results of our efforts to engineer the production of TAL suggest that oleaginous organisms may be better defined, not by the accumulation high levels of lipids, but by the high carbon fluxes into acetyl CoA synthesis under nitrogen limiting conditions. In unengineered cells, this acetyl CoA is directed into fatty acid synthesis, but we show that this acetyl CoA can be directed into other heterologous pathways. Finally, oleaginous organisms are typically classified into a single broad stroke mostly centered on the ACL pathway: leverage the unique ACL pathway to manipulate carbon flux into fatty acid synthesis. However, the findings reported in this work identify the contribution of beta oxidation to TAL synthesis. Instead of focusing on just the ACL pathway, the results of this work warrant further research into the unique metabolic subtleties of each oleaginous organism.

CHAPTER 7: Materials and Methods

Table 7.1 Strains and plasmids utilized in this work

Strain Name	Genotype
ySBS1733	CLIB89, ATCC 20460, mating type A
yJY1832	ySBS1733, *ura3Δ*
yJY1853	yJY1832, *gut2Δ::URA3*
yJY1862	yJY1832, *ku70Δ::URA3*
yJY1886	*yJY1832 ura3Δ, URA3::pol4Δ*
yJY1900	*ura3Δ, leu2Δ, pol4Δ*
yJY1948	ySBS1733, *ura3Δ*
yJY1953	yJY1948, *leu2Δ-1 ura3Δ*
yJY2002	yJY1948, *leu2Δ::URA3 ura3Δ*
yJY2006	yJY2002, *leu2Δ ura3Δ*
yJY1984	yJY1953, *leu2Δ-1 mae1Δ::URA3*
yEJ1986	yJY1953, *leu2Δ-1 rpd3Δ::URA3*
yEJ1988	yJY1953, *leu2Δ-1 utr1Δ::URA3*
yJL2008	yJY2006, *leu2Δ ura3Δ pex10*
yJL2011	yJY2006, *leu2Δ ura3Δ gpd1*
yJY2039	yJY1948, with integrated copies of 2-PS
yJY2042	yJY2006, *leu2Δ ura3Δ* Tn insertion upstream of YALI1_B21493 (*PRB1*)
yJY2043	yJY2006, *leu2Δ ura3Δ* Tn insertion location unknown
yJY2044	yJY2006, *leu2Δ ura3Δ* Tn insertion upstream of YALI1_F11261
yJY2045	yJY2006, *leu2Δ ura3Δ* Tn insertion upstream of YALI1_E18713
yJY2046	yJY2006, *leu2Δ ura3Δ* Tn insertion upstream of YALI1_C14402r (*5S*)

Plasmid Name	Description
pJY3731	*TEF1* promoter driving hrGFP, *URA3*
pJY3830	*TEF1* promoter driving hrGFP, *LEU2*
pPS3911	TIR-flanked *LEU2*
pJY3919	*TEF1* promoter driving recoded Hermes transposase; TIR-flanked *LEU2*
pMT3928	*TEF1* promoter driving recoded Hermes transposase
pDTN3950	Template for *URA3* cassette for *RPD3*, *UTR1*, and *MAE1*, *KU70* KO cassettes.
pDTN3997	3x *TEF1* intron promoter driving 2-PS
pSBS4053	CRISPR plasmid for *Y. lipolytica*
pJL4064	CRISPR plasmid to disrupt *SNF1*
pJY4068	*SNF1* expressed on a plasmid with mutant PAM site for plasmid

	shuffle
pJL4070	CRISPR plasmid to disrupt *GPD1*
pJL4071	CRISPR plasmid to disrupt *HIS3*
pJL4072	*HIS3* expressed on a plasmid with mutant PAM site for plasmid shuffle
pJL4074	CRISPR plasmid to disrupt *PEX10*
pJY4089	*TEF1* promoter driving recoded Hermes transposase; TIR flanked 2-PS

Materials and Methods described in Chapter 2:

Yeast and Bacterial Strains

Escherichia coli DH5α was grown at 37°C in Lysogeny Broth (LB) media with 100 μg/ml ampicillin to select for plasmid retention. *Y. lipolytica* wild-type (wt) strain CLIB89 was obtained from ATCC (20460). All strains utilized in this study were derived from this strain (Table 1). Yeast were cultured at room temperature (rt) (22°C to 25°C). Cultures were grown in yeast peptone dextrose (YPD) [1% yeast extract (wt/vol), 2% peptone (wt/vol), and 2% dextrose (wt/vol)], minimal medium [(0.67% yeast nitrogen base (YNB, Difco), 2% (wt/vol) dextrose,0.5% (wt/vol) ammonium sulfate)], or synthetic complete (SC) media lacking the indicated nutrients (Guthrie and Fink 1991). A uracil auxotroph (yJY1832) lacking the *URA3 gene* (currrently annotated as YALI0E26741g) was generated by targeted integrative transformation with a cassette generated by PCR of the sequence from -484 to -1184 upstream of the *URA3* ORF and + 598 bp to 1787 bp downstream. In addition to eliminating the open reading frame (ORF), this eliminated sequences present in the *URA3* marker used in subsequent manipulations. Targeted integration would therefore generate a strain with a 1925 bp deletion, including the 855 bp *URA3* ORF. The transformation mix was plated on 5-fluoroorotic acid (5-FOA) to select for *ura3*Δ isolates. Strains with targeted deletions were confirmed by polymerase chain reaction (PCR) using genomic DNA and primers located at -1200 bp upstream and 1822 bp downstream of the URA3 coding region. This produced an amplicon that was roughly 2000 bp smaller than the product using genomic DNA from a wt strain. Subsequent Sanger sequencing (Genewiz Inc.) confirmed the exact 1925 bp deletion.

Gene knockouts utilizing the *ura3*Δ background were constructed with a PCR-generated knockout cassette containing the *URA3* gene as a selection marker. Primer numbers and sequences are listed in Table S2. The cassette consisted of the *URA3* promoter, ORF, and downstream sequences lacking in the deletion strain, flanked by homologous 5' and 3' regions flanking the target gene. The homologous flanking sequences utilized for specific deletion cassettes are provided in Table S1. Target genes included *LEU2* (YALI0C00407g), *YKU70* (YALI0C08701g) and *POL4* (YALI0D12364g). All *leu2*Δ strains lack the *LEU2* ORF and 878 bp of 5' upstream sequence and 44 bp of 3' downstream sequence. The *ku70*Δ strains lack only the *KU70* ORF to mirror construction of a previously published strain (Kretzschmar, Otto et al. 2013). The *pol4*Δ strain is lacking the *POL4* ORF along with 375 bp of upstream of the coding region and 481 bp of downstream of the coding region. Yeast strains utilized in the study are listed in Table 1.

Plasmid Construction

The *Y. lipolytica URA3*-marked and *LEU2*-marked *ars18* plasmids were derived from the bacterial AmpR plasmid pUC18 (Yanisch-Perron, Vieira et al. 1985) (Figure S1) . The *ars18* (GenBank accession number AX023557) (Fournier, Guyaneux et al. 1991), *URA3* (pJY3731) and *LEU2* (YALI0C00407g, for pJY3830) sequences were generated by PCR from *Y. lipolytica* genomic DNA The *URA3* marker consisted of the region 590 bp upstream and the region 580 bp downstream flanking the *URA3* ORF. The *LEU2* marker consisted of the regions 763 bp upstream and 751 bp downstream of the *LEU2* ORF. The *ars18* sequence was cloned into the pre-existing *Eco*RI site of the pUC18 backbone. Afterwards, pUC18 was modified via DpnI mutagenesis to add a novel NotI restriction enzyme site (Zheng, Baumann et al. 2004). Selection markers were cloned into this NotI restriction site. Complete sequence information and maps for plasmids are provided in Figure S1 and S2.

Plasmids used to generate linear DNA cassettes for gene knockout/in were created as follows. Portions of the regions immediately upstream and downstream of the coding sequence, used for homologous targeting (exact lengths listed in Table S1), of the gene of interest were synthesized by PCR from genomic DNA. These sequences were fused together by PCR using the strand-overlap elongation (SOEing) procedure with a SwaI restriction enzyme site inserted in the primer-derived overlap between the two fragments (Horton, Hunt et al. 1989). The resulting fusion was inserted into the TOPO vector, pCR2.1 (Invitrogen). *URA3* was amplified from genomic DNA using primers containing SwaI restriction sites (including the ORF and 700 bp upstream and 1189 bp downstream) and was inserted between these flanking sequences by SwaI digestion of the plasmid and ligation into the PCR-amplified SwaI-digested *URA3* cassette into the plasmid backbone. The resulting plasmid was used as the PCR template for knockout/in cassette synthesis. Plasmids utilized in the study are listed in Table 2. Primer sequences are described in Table S2.

Transformation Protocols

For all transformations, 0.1μg (0.018 ρmole) plasmid DNA or 0.1ρmol linear DNA was added to the competent cells.

Protocol #1 (Barth 1996; Kretzschmar, Otto et al. 2013) – *Y. lipolytica* was grown to OD_{600} = 1.5-2.0 in YPD, pelleted by centrifugation and resuspended in 1/10 volume of lithium acetate (LiOAc) solution (100mM LiOAc, 100mM Tris-EDTA pH7.4). The cells at a concentration of 10 OD/ml were placed on a nutator for between 1 to 4 h. Cells were centrifuged and resuspended in 1/100 the original volume in LioAc solution, and 10 OD cells (in 100µl) were used for each transformation. Fifty microliters of salmon sperm DNA (10mg/ml) and transforming DNA were added to cells. Following a 15 m incubation at room temperature (rt), 300µl of a polyethylene glycol (PEG) solution (40% PEG-3350, 100mM LiOAc, 100mM Tris-EDTA pH7.4) was added, followed by a 1 h rt incubation. The solution was subjected to a 10 m heat shock at 39°C and the transformation solution was plated on the appropriate solid medium.

Protocol #2 (Chen, Beckerich et al. 1997) – Competent cells were produced as stated above. Cells were pelleted and 10 OD cells were used for each transformation. These cells were resuspended in 200µl One Step Buffer (45% PEG4000, 100mM LiOAc pH 6.0, 0.25mg/ml salmon sperm DNA, 100mM DTT) and transforming DNA. The solution was subjected to a 1 h heat shock at 39°C and the transformation solution was plated on the appropriate solid medium.

Protocol #3 (Gietz and Schiestl 2007) – Competent cells were produced as stated above. Cells were pelleted and 10 OD cells were used for each transformation. These cells were resuspended in 350µl Transformation Mix solution (33% PEG-3500, 100mM LiOAc pH6.0, 0.25mg/ml salmon sperm DNA) and transforming DNA. The solution was subjected to a 40 m heat shock at 39°C and the transformation solution was plated on the appropriate solid medium.

Dilution Assays

Wild-type (wt) (CLIB89), *ku70*∆ (yJY1862) and *pol4*∆ (yJY1886) strains were grown in YPD overnight and diluted the following morning. Cells were harvested in log phase at an OD_{600} of 1 to 2 and diluted to 1.0 in YPD. Aliquots of each strain were subjected to 39°C heat shocks of 0, 30 or 60 m duration. Serial dilutions were performed with each aliquot to generate 1:10, 1:100 and 1:1000 dilutions. Five microliters of the original aliquot and each dilution were spotted on YPD or minimal medium. Plates were incubated at rt, 30°C or 34°C for 24 h and photographed.

Materials and Methods described in Chapter 3:

CLIB89 and CLIB122 strain relationship

Y. lipolytica strains in industrial use are descended from strains isolated from diverse ecological niches and in some cases, crossed with other isolates and sporulated to yield new haploid strains. One strain with significant genomic sequence available is W29, which was isolated from waste water in Paris, France. It is designated in the Collection de Levures d'Interet Biotechnologique (CLIB) as CLIB89. A second is E150, which is a haploid derivative of a cross of CBS6124-2 isolated from an American corn processing plant and CLIB89. This strain was designated CLIB122. The CLIB122 genomic sequence is the most thoroughly characterized *Y. lipolytica* genomic sequence (http://www.ncbi.nlm.nih.gov/genome/genomes/194?). In order to have a comparably detailed assembly of the progenitor strain CLIB89, haploid *Y. lipolytica* CLIB89 was purchased from ATCC (ATCC 20460) and the genome sequence was determined.

Culture conditions

For genomic sequencing, CLIB89 was cultured in 2% Yeast extract-1% Peptone 2%-Dextrose (YPD) (Sherman 2002). For Illumina sequencing DNA was extracted according to standard protocols; RNA was removed by RNase digestion (Ausubel, Brent et al. 2007). DNA was sheared to appropriate length using the S2 Ultrasonicator (Covaris). Libraries were prepared for sequencing using the NEXTflex Rapid DNA-Seq Kit (Bioo Scientific). For PacBio sequencing, cells were spheroplasted by treatment with Zymolyase (Seikagaku Corporation). Spheroplasts were pelleted, lysed and digested with RNaseA (Fermentas) and Proteinase K (Fisher Scientific). DNA was isolated and then eluted from the Qiagen Anion-Exchange Resin. For datasets YLP13 and YLP14, corresponding to PacBio RS sequencing, high MW DNA was extracted with the Genomic-Tip 20/G kit (Qiagen). DNA was fractionated to 4-50 kb using a BluePippin pulsed-field gel electrophoresis system (Sage Sciences). PacBio SMRT Bell sequencing libraries were prepared using the manufacturer's DNA SMRT kit. Results of RNA sequencing will be published elsewhere. Transcripts from cells grown under several conditions were combined in order to maximize the potential of transcriptomics to identify reading frames. For RNA sequencing cells were lysed, and RNA was processed into KAPA stranded libraries for Illumina PE100 sequencing according to manufacturer's instructions.

Next-generation Illumina and PacBio sequencing

For the genome assembly described below (Implementation), four DNA sequencing datasets were generated from a combination of Illumina HiSeq 2500 and PacBio RS II reads (Table 1). Datasets YL97B and YL110, corresponded to Illumina single-read (SR) 97 cycles and paired-end (PE) rapid run 110 cycles, respectively. For YLP13 and

YLP14 sequencing was conducted on eight and four SMRT cells, respectively. Longer inserts were prepared for YLP14 (Table 1). Illumina sequencing data were processed and demultiplexed using CASAVA1.8.2. PacBio RS II was processed using SMRT Portal.

***De novo* genome assembly**

High-quality short Illumina reads and long PacBio RS II reads were combined in a custom hybrid approach to assemble the CLIB89 genome (Table 1). Details are provided in S1 Text.

Contig assembly of Illumina short reads

Sequences from YL97B and YL110 Illumina datasets were assembled into contigs and filtered as described in S1 Text, Materials and Methods using the Velvet *de novo* assembler (Zerbino and Birney 2008).

Extending and scaffolding of Illumina contigs with long PacBio reads

PacBio long reads from datasets YLP13 and YLP14 were used to scaffold the *de novo* contigs generated from Illumina analysis. This PacBio scaffolding allowed the determination of both immediate mate(s) to each contig, as well as long distance information up to four mates away. This information was used to determine the placement of contigs spanning repetitive regions in the genome. The PacBio scaffolding was accomplished through the following four steps: i) matching PacBio long reads with the first-stage contigs via BLAST (Altschul, Gish et al. 1990) and extending each BLAST hit via a combination of in-house software and ClustalW (Larkin, Blackshields et al. 2007) to generate mapped sequences with computed percent identity; ii) filtering to remove unreliable mapped sequences, sequences that will not provide scaffolding information, and sequences leading to ambiguous or conflicting cases; iii) selecting the closest, non-repetitive right-mate contig and incorporating that into the growing assembly; and iv) identifying the correct mate for a repeated contig via long-distance contig information.

Assembling junctions between extended contigs

Successful scaffolding of contigs from the previous stage enabled the identification of junction sequences. Moreover, the consensus junction sequences resulting from the multiple sequence alignment was estimated to have an error rate of 8%, significantly lower than the error rate of a single PacBio read.

The junction sequences were obtained via ClustalW by generating a multiple sequence alignment (MSA) of PacBio reads from datasets YLP13 and YLP14 spanning each junction region plus 200 bp into each flanking de novo contig. Positions in the MSA were selected for the consensus sequence only when: i) enough reads confirmed its existence, and ii) the reads had high agreement, where the two parameters values (i, ii) were optimized for each junction for which the target consensus sequence was known. These junction sequences were further corrected for consensus by utilizing previously unused reads from datasets YL97B and YL110, where unaligned single reads in YL97B with respect to the contigs were reused by Velvet to generate additional comparison contigs, and paired reads in YL110 with part of a read in the junction regions were extracted with Eland v2e. The new contigs and selected paired reads were used iteratively to complete the error-correction. The resulting scaffolds were assigned chromosomal names based on comparison to chromosome designations in the CLIB122 assembly.

Bridging the gap in chromosome C and extending chromosome ends

From the aforementioned assembly pipeline, two scaffolds matched with chromosome Yali0C of the CLIB122 assembly. The junction sequence was assembled by: i) aligning paired-reads in YL110 iteratively to the scaffold ends to extend the scaffold sequences; ii) extending contigs by aligning unused PacBio reads at each iteration of scaffold extension via the same protocol for PacBio reads described in the previous two stages; iii) stopping the iterations when enough PacBio reads were matched to both scaffold ends. The same approach was used to extend the chromosome ends until no more bases could be added. The chromosome C contig joint was confirmed to overlap by PCR utilizing primers annealed to unique sequences at the ends of the two contigs (S1 Table).

Mapping of chromosomal ends

Of the twelve chromosomal ends, two were similar between the CLIB89 and CLIB122 assemblies; five were longer in CLIB122. Because these five CLIB89 sequences overlapped they were extended into the CLIB122 assembly (Materials and Methods S1 Supplemental Materials and Methods).

In an effort to extend the chromosomes as close to the telomeres as possible, five CLIB89 terminal sequences were extended by joining termini of CLIB89 contigs to overlapping CLIB122 sequences and these joints were verified by existing overlapping long reads. CLIB122 sequence was appended to the ends of the CLIB89 assembly as follows: Yali0B 3' end position 3,044,622 (387 bp); Yali0C 5' end 1 8,912 bp; Yali0C 3' end position 3,353,699 (12,578 bp); Yali0D 5' end 1 3,177 bp; Yali0E 3' end position 4,188,128 bp (10,433 bp); Yali0F 3' end position 3,999,287 bp (3,679 bp). PCR amplification utilizing primers 4956/4957, 4958/4959, and 4960/4961 were used to confirm these additions to the 5’ end of Yali1C, the 3’ end of Yali1B, and the 5’ end of Yali 1D respectively. A flanking primer in the CLIB122 sequence together with a rDNA primer was used to confirm the position of one rDNA repeat. Because the *Y. lipolytica* rDNA genes occur as tandem repeats in multiple clusters, a complete rDNA sequence and non-transcribed spacer could be generated using outward priming oligonucleotides based on a single fragment of rDNA sequence in CLIB122 (S3 Table). This demonstrated the occurrence of rDNA sequence on chromosome F. However, there are multiple clusters of rDNA reported in *Y. lipolytica* (Gaillardin, Mekouar et al. 2013) and it is therefore not possible to determine that the intergenic sequence amplified by PCR was derived from chromosome F.

Irys long-range mapping analysis

Irys long-range genome mapping (BioNano Genomics, Inc.) enables comparison of chromosomal restriction maps to chromosomal sequence for genome sequence validation and genome comparisons. In our application DNA molecules of average length 285 kb were subjected to nicking with a single-strand-specific restriction endonuclease followed by nick translation to introduce fluorescent tags. DNA was stained and imaged during low-voltage electrophoresis in Irys instrument nanochannels.

The raw image data were converted to digital representations of the restriction site-specific labeling and the resulting tag patterns of 621,169 molecules were assembled *de novo* into 31 contigs using IrysSolve software (Fig 3.1A, Table 3). These contig tag patterns were aligned to a virtual restriction tag pattern generated from the hybrid Illumina-PacBio YALI1 sequence. This alignment showed overall agreement between the two assemblies. However, the length of the Irys assembly was 25.246 Mb, significantly longer than either CLIB89 or CLIB122 sequence assemblies. This difference could be explained by heterogeneity of the lengths of chromosomal ends in which Irys mapping was biased for the longest extensions or redundancy in mapping terminal repeated regions. Consistent with these possibilities, inspection of the Irys assembly in telomeric regions showed that Irys contigs at the left end of chromosomes A, C, and F and the right end of chromosome B were each nearly 1 Mb longer than the corresponding YALI1 chromosomes and contained a distinctive ~10 kb unit repeat. Analysis of these regions is discussed further below.

DNA extraction and labeling was according to BioNano Genomics protocols. Briefly, CLIB89 cells were spheroplasted by treatment with Zymolyase (Seikagaku Corporation), immobilized in low-melting point agarose matrix, and treated with proteinase K (Qiagen), washed with TE, digested with RNAse and washed again with TE. Agarose was melted and digested with GELase (Epicentre Biotechnologies) to recover genomic DNA. Buffers were exchanged by drop dialysis and DNA was stained according to QuBit dsDNA HS kit instructions, sonicated and quantified on a Qubit Fluorometer. DNA molecules of average length 285 kb were subjected to digestion to completion with single-strand nickase Nt.BspQI. Finally 300 ng of DNA was labeled by limited-drive nick translation in the presence of a fluorophore-labeled nucleotide. Labeled nicks were repaired using a thermostable polymerase and ligase.

Eight microliters of DNA at a concentration of 4.6 ng/microliter DNA was loaded into a flowcell of the IrysChip for imaging in the Irys instrument in a low voltage electric field controlling sample flow of individual molecules through nanochannels for 30 cycles.

Under laser excitation images of DNA were captured by EM-CCD in the Irys Instrument and using AutoDetect software processed into digitized molecule image files. IrysView software was used to analyze and visualize these data. IrysSolve running on an independent server was used to assemble the *Y. lipolytica* CLIB89 genome. This analysis generated 31 contigs. These contigs were aligned with YALI1 sequence at a confidence threshold of negative log of P value =45. Chromosomes A, B, C, and F showed substantial extensions relative to the hybrid sequence assembly. The extended regions displayed striking tandem unit repeats of ~10kb tagged at intervals consistent with the Nt.BspQI digest pattern predicted for rDNA-coding sequence (S4 Table).

Gene identification using RefSeq/NCBI *Y. lipolytica*

CLIB122 gene sequences available in RefSeq Yl were compared to the CLIB89 assembly via a combination of various BLAST (Altschul, Gish et al. 1990) and Exonerate (Slater and Birney 2005) alignment algorithms to maximize consistency with previously identified genes. To find the equivalent features across the assemblies, Exonerate was used to map CLIB122 RefSeq Yl features by their respective nucleotide sequences, both with and without flanking nucleotide sequences, to the CLIB89 assembly. The Exonerate mapping with flanking sequences identified the equivalent genes in the two assemblies, while mapping without flanking sequences helped to identify additional paralogs present in CLIB89. The EST2GENOME alignment algorithm was used on protein-coding sequences to take into account intron variation, while AFFINE:LOCAL (a local alignment similar to the Smith-Waterman-Gotoh algorithm) was used on non-coding sequences. Loci mapped in CLIB89 were assigned corresponding RefSeq Yl annotations.

Gene identification using YGAP

YGAP is an automated yeast/fungal genome annotation services available online at http://wolfe.ucd.ie/annotation/. CLIB89 and CLIB122 assemblies were analyzed in parallel with the seven chromosome sequences of each assembly used as “Scaffolds” in creating a new YGAP project. In addition to identification of coding sequences, YGAP identifies tRNA genes (tDNAs) using tRNAscan-SE (Lowe and Eddy 1997) and excludes them from coding regions, and identifies Ty LTR retrotransposons. For a parallel comparison, CLIB122 YALI0 was analyzed using YGAP as well. The YGAP CLIB89 gene set included 6,448 loci, 5,938 protein-coding sequences, and 510 tDNA (chromosomal and mitochondrial), compared to YGAP CLIB122 gene set of 6,467 loci with 5,930 protein-coding sequences, and 537 tDNA (including mitochondria and chromosomal).

Gene refinement and validation using SnowyOwl HMM

The SnowyOwl pipeline is based on ranking models generated by various HMM gene predictors (Reid, O'Toole et al. 2014). To run the SnowyOwl pipeline for the present study, RNA-Seq reads were mapped and assembled into *de novo* transcripts via TopHat and Cufflinks (Trapnell, Pachter et al. 2009), which was then used as the initial transcript model to help train the subsequent SnowyOwl pipeline. Cufflinks-assembled transcripts were aligned to the CLIB89 assembly using Tophat2, and then used to generate initial models for training a Hidden Markov Model gene predictor, together with the intron and transcribed positions revealed by the Tophat2 mappings. Next, Genemark-ES (Lomsadze, Ter-Hovhannisyan et al. 2005; Ter-Hovhannisyan, Lomsadze et al. 2008) was run on the CLIB89 assembly to generate GeneMark gene models, which were ranked based on RNA-seq data to produce a set of high-confidence gene models as the training set for another *ab initio* gene predictor, AUGUSTUS (Stanke and Morgenstern 2005). AUGUSTUS was run several times with different parameter settings and generated models scored using homology-based evidence, exon-intron boundaries and coverage. Highly-scored models with the best agreement with RNA-Seq data and homology evidence were preserved by default; imperfect models that lacked similarly high RNA-Seq coverage were also captured and added to the list, but were flagged as imperfect. The homology evidence used included BLASTx sequence homology search against the NCBI fungal databases and Uniprot protein database (UniProt 2015). In total, from the HMM models, SnowyOwl predicted 7,482 protein-coding sequences over 100 codons in length on the CLIB89 assembly—5,464 of which corresponded to RNASeq transcripts. Of the 7,482 SnowyOwl protein-coding sequences, 6,311 corresponded to protein-coding sequences identified by either YGAP or RefSeqYl, leaving a total of 1,171 additional protein-coding sequences identified uniquely by SnowyOwl (Fig 3.2).

Final merged set of annotated genes

Existing locus/gene names from CLIB122 YALI0 were reassigned as locus/gene names in the independent CLIB89 YALI1 assembly (S2 Table). UniProt IDs present in the RefSeq Yl were retained in the annotations. The RefSeq mapped sequences, YGAP genes, and SnowyOwl protein coding sequences used common coordinates based on the CLIB89 DNA-seq assembly. The output files of the three pipelines were merged together via the LocusPocus locus-based annotation script in the Aegean Toolkit (Duvick, Standage et al. 2016) (http://standage.github.io/AEGeAn). Further processing based on coding sequence length and integrity were performed using in-house scripts that then generates both IGV ready gff3 file and the TBL file for NCBI submission. The TBL undergoes quality checks through NCBI's Sequin annotation upload service (http://www.ncbi.nlm.nih.gov/projects/Sequin/download/seq_download.html), where problematic gene features are resolved through iterative manual curation.

Consensus defined solo LTR identification

Genome sequences for *Y. lipolytica* CLIB89 in FASTA format were obtained from hybrid assembly of Illumina and PacBio reads. For analysis of CLIB122 Genolevures database sequences were used [(previously) http://www.genolevures.org/index.html#; (currently) Genome Resources for Yeast Chromosomes website (http://gryc.inra.fr) and http://www.ncbi.nlm.nih.gov/genome/genomes/194?].http://www.ncbi.nlm.nih.gov/genome/genomes/194?].http://www.ncbi.nlm.nih.gov/genome/genomes/194?].http://www.ncbi.nlm.nih.gov/genome/genomes/194?].http://www.ncbi.nlm.nih.gov/genome/genomes/194?].http://www.ncbi.nlm.nih.gov/genome/genomes/194?].http://www.ncbi.nlm.nih.gov/genome/genomes/194?].http://www.ncbi.nlm.nih.gov/genome/genomes/194?].

Allelic genomic features and variants analysis

Genome features were extracted from both CLIB122 YALI0 and CLIB89 YALI1 assemblies. For both lists of genome features, we used the locuspocus program of the AEGeAn toolkit (Standage 2010-2015) to designate both annotated genomic features and intergenic regions as separate interval loci (iLoci). A two-layered Exonerate alignment (first by using CLIB122 iLoci as query and CLIB89 iLoci as targets and second by vice versa) was used to determine allelic iLoci between the two assemblies, as well as to quantify the number of mismatches and indels between the allelic iLoci

(Table S8). In our analysis, nucleotides present in CLIB89 and not CLIB122 was classified as an insertion and vice versa. Variants density is reported as number of Mismatches/insertion/deletions per kb.

Transposable elements of CLIB89 YALI1 (reported in Table S7 and S8) were mapped to the CIRCOS diagram according to their genomic coordinates. To map the transposable elements of CLIB122 YALI0, two methods were utilized. CLIB122 Elements that are allelic were mapped to the same chromosomal location as the CLIB89 counterpart. For CLIB122 elements that were not allelic, flanking iLoci were then utilized to assign a corresponding genomic location.

Mitochondrial genome analysis

The mitochondrial genome of Y. lipolytica was previously annotated and reported (Kerscher, Durstewitz et al. 2001; Gaillardin, Neuveglise et al. 2012). Mitochondrial genome analysis was performed utilizing RNAweasel (http://megasun.bch.umontreal.ca/RNAweasel/)**(Lang)**, an automated annotation program for organellar genomes with intron detection and other features tuned for mitochondrial genomes with manual adjustments. Variants analysis with CLIB122 mitochondrial assembly was performed using the SNP detection pipeline of MUMmer (http://mummer.sourceforge.net/) by using CLIB122 as query and CLIB89 as target to be consistent with the chromosomal variants analysis.

Materials and methods described in Chapter 4

Strains and plasmids and growth conditions

Escherichia coli DH5 was grown at 37°C in Lysogeny Broth (LB) medium supplemented with 100 µg/ml ampicillin to select for plasmid retention. *Y. lipolytica* wt strain CLIB89 was obtained from ATCC (20460). All yeast strains used were isogenic to CLIB89. To facilitate transposon mutagenesis and other investigations in this work, strains were modified and are listed in Table S1. Yeast were cultured at room temperature (22°C to 25°C) or at 28ºC as indicated. Cultures were grown in YPD [1% yeast extract (wt/vol), 2% peptone (wt/vol), and 2% dextrose (wt/vol)], YPG [1% yeast extract (wt/vol), 2% peptone (wt/vol), and 2% glycerol (wt/vol)], SD -Leu (nitrogen-enriched) medium [6.7g/L yeast nitrogen base, 2% dextrose, inositol, adenine sulfate, and complete amino acids lacking leucine], or nitrogen-depleted medium [0.625 g/L $(NH_4)_2SO_4$, 5 g/L Na_2SO_4, 2% dextrose, 1.7 g/L yeast nitrogen base without amino acids and $(NH_4)_2SO_4$, and 35 mg/L uracil] (Guthrie and Fink 1991).

Y. lipolytica strains, plasmids and primers used in this study are summarized in Table S1. A *URA3*-marked CEN/ARS18 donor plasmid (pJY3919) was constructed for Hermes transposon mutagenesis. This donor plasmid contains a selectable cassette that is mobilized by codon-optimized Hermes transposase expressed under the *TEF1* promoter (Muller, Sandal et al. 1998; Tai and Stephanopoulos 2013). The cassette consists of *LEU2* (beta-isopropyl malate dehydrogenase) flanked by the Hermes terminal inverted repeats (TIRs). Two negative control plasmids were constructed based on the same vector. One contains TIR-flanked *LEU2* and lacks the transposase (pPS3911), while the other contains the transposase and lacks TIRs on either side of the *LEU2* marker (pMT3928) (Table S1).

Yeast transformations were performed using a modified lithium acetate transformation protocol (Chen, Beckerich et al. 1997; Gietz and Schiestl 2007). Strains were inoculated into a 3 mL culture of YPD, grown overnight at 28°C, diluted 1:50 into fresh YPD and harvested at 1.0-1.5 OD_{600}. Cell pellets were resuspended in 1x TE/LiOAc (10 mM Tris, 1 mM EDTA, 100 mM lithium acetate) at a ratio of 250 µL per 0.5 OD_{600}. Transformation reactions contained 300 µL of 1x TE/LiOAc/DMSO/PEG40 (TE/LiOAc plus 10% dimethyl sulfoxide, and 40% PEG 3350), 500-600 ng of plasmid DNA, and 50 µL of cell resuspension. Transformations were incubated at 30°C with rotation for 3 hrs, heat shocked at 39°C for 30 min, pelleted and resuspended in 200 µL water and plated onto the appropriate selection medium.

Hermes transposition in *Y. lipolytica*

For Hermes transpositions, uracil and leucine double auxotroph strains were transformed with pJY3919 or negative control plasmids pPS3911 and pMT3928. Test experiments indicated that 54.7%±6.7% of pJY3919 transformants had transposition events while pPS3911 or pMT3928 transformants had a negligible frequency of transposition. Analysis of *LEU2* copy number showed that only one in 28 tested strains had more than a single copy of *LEU2* (Fig. S1). For genome profiling, yJY1953 (Trial 1) and yJY2006 (Trial 2), were transformed with pJY3919. The transformation protocol was followed as described above with an additional modification: after heat shock, pellets were resuspended in 1 mL water and then plated onto SD-Leu medium supplemented with 5-fluoroorotic acid (5FOA) (Boeke, Trueheart et al. 1987) in 150 mm plates and incubated at room temperature for 2 days until colonies appeared. Room temperature was chosen as a relatively permissive condition, recognizing that growth at the temperature of maximum growth rate (28-30ºC) would likely cause more genes to be classified as essential. Approximately 2.2 million colonies were harvested and re-plated onto SD –Leu +5FOA medium to select for cells containing the *LEU2*-tagged transposon that no longer contained the *URA3*-tagged pJY3919 backbone. 5FOA-resistant colonies were pooled to represent the Gen0 population. For competitive growth experiments in YPD or YPG, 4 x 100 mL replicate cultures were inoculated with Gen0 cells with a starting OD_{600} = 0.05. For passaging, every 12 hrs of growth at 23°C, cultures were re-inoculated to OD_{600} =0.05 in fresh medium. At each passage the number of generations was estimated based on the $OD_{600,}$ and at roughly 20 and 80 generations, 100 OD_{600} cell pellet from each replicate was harvested for genomic DNA isolation and processing. For simplicity these are referred to as Gen20 and Gen80.

Sequencing library preparation

The gDNA was isolated as described above (Amberg, Burke et al. 2005). After extraction, gDNA was treated with DNase-free RNaseA (100 µg/mL) (Sigma- Aldrich) at 37° C for 4 hrs, followed by Pronase (250 µg/mL) (Sigma-Aldrich) at 37°C for 2 hrs. After both RNase and Pronase treatment, another phenol:chloroform extraction was performed. Approximately 30 µg of this treated DNA was sheared to less than 1 kb in a 100 µl volume using the Biorupter Standard (Diagenode). DNA size was confirmed by gel electrophoresis. Sheared DNA was concentrated using the Clean and Concentrator with 5X ChIP DNA Binding Buffer (Zymo Research). DNA ends were made blunt using the Fast DNA End Repair kit (Thermo Scientific); 3' adenylated using the Klenow Exo-enzyme with 0.2 mM dATP (Fisher Scientific). Custom sequencing adapters (0.5 µM) with Illumina multiplex barcodes (Table S1) were ligated to DNAs in each sample using

T4 ligase (Thermo Scientific) by incubating at 22°C for 12 hrs followed by 65°C for 10 min. Between enzyme reactions, samples were purified using the Agencourt AMPure XP purification beads (Beckman Coulter, Inc.). Hermes TIR –gDNA borders were enriched in PCR using the high-fidelity KOD polymerase (EMD Millipore) as follows: 95°C 5 min, 5 cycles of 95°C 30 s, 65°C 30 s, and 70°C 30 s, followed by 20 cycles of 95°C 30 s, 61°C 30 s, and 70°C 30 s, and a final 70°C incubation for 2 min. Forward primer KP5113 anneals 26 bp from the downstream 3'-end of the TIR, and reverse primer KP5112 anneals to the adapter upstream of the Illumina multiplex barcode. The same PCR primers were used for all samples. To avoid PCR bias each library sample was split into 16 technical replicates, then recombined and size selected using the AMPure XP Beads. Each PCR reaction contained 25 ngs of template DNA. After sizing/purification of PCR products samples were submitted to the UCI Genomics High-throughput Facility for Bioanalyzer 2100 (Agilent) analysis, KAPA quantification and single-end 100 cycle sequencing on the Illumina HiSeq 2500.

Sequencing data analysis

Briefly, reads were first trimmed to remove the random 5-nt sequence, then sequences downstream of KP5113 primer binding site but not matching the expected last 26 nts of the Hermes TIR and representing mis-primed reads were filtered out. Passing DNA sequences were trimmed to remove the 26 nt TIR sequence and mapped to the CLIB89 YALI1 genome (Magnan, Yu et al. 2016) using Bowtie2 (Langmead and Salzberg 2012). Each of the 534,589 unique reads mapped defined a Hermes insertion site; non-unique reads mapping to multiple positions, such as those representing insertions into tDNAs and transposons, were randomly assigned (less than 1.5% of reads). All raw

Hermes insertion site positions and nearest annotated gene are provided in Table S2. Hermes insertion coordinates were mapped to gene annotations using BEDTools (Quinlan and Hall 2010). Essential gene classification was done using custom Pythonv2.7 scripts which are available upon request (see below). Insertion site motif analysis was done using WebLogo (Schneider and Stephens 1990; Crooks, Hon et al. 2004). Phenotype data from *S. cerevisiae* for comparison to *Y. lipolytica* was acquired from the *Saccharomyces* Genome Database (SGD)(Cherry 2016). *Y. lipolytica* homologs to *S. cerevisiae* genes were acquired from the CLIB89 genome annotation (Magnan, Yu et al. 2016). For simulated transposition, custom scripts were used to scan the CLIB89 genome for all 5’nTnnnnAn3’ or 5’nAnnnnTn3’ sites, and these sites were mapped to the annotated CLIB89 genome using BEDTools. All graphs were prepared in R Studio (2016).

Essential gene classification

Mutants were analyzed for genes with a statistically significant underrepresentation of Hermes insertions relative to an expected number of hits based on gene length (L). Similar to Gerdes et al. (Gerdes, Scholle et al. 2003), our gene classification system (GCS) uses the Poisson distribution to test the probability of observing k insertions assuming a rate of 97.5 bp/hit for genes (λ) and a rate of 24.37 bp/hit for intergenic regions (λ^i). When the probability (P) of witnessing k insertions in a gene of length L was below 0.05, the gene was classified as essential.

$$P_k(L) = \frac{[\left(\frac{L}{\lambda}\right)*R]^k}{k!} * \ e^{-[\left(\frac{L}{\lambda}\right)*R]}$$

Local features flanking a gene such as heterochromatin or gene density can impact the hit frequency; thus for the local hit density a 5’ and 3’ 1-kb flanking region (L_f) was used as a metric for determining the local correction ratio (R). R is defined as the ratio of observed local hit frequency over the expected. Observed local hits includes the sum gene hits (H_{gene}) and each 1-kb flank (H_{flank}); whereas expected hits includes the sum of expected gene hits $\left(\frac{L}{\lambda}\right)$ and flanking $\left(\frac{L_f}{\lambda_i}\right)$. Thus R is defined as

$$R = \frac{(H_{gene} + H_{flank})}{\left[\left(\frac{L}{\lambda}\right) + \left(\frac{L_f}{\lambda_i}\right)\right]}$$

When R < 1, it was multiplied by the expected gene hit value in the Poisson equation above; however, when R ≥ 1 no correction was applied (i.e. R=1).

In calculation of the GCS only unique insertions were included. Two classes of hits were excluded from the GCS calculation. First, hits in several genes were restricted to introns, consistent with tolerance for intronic, but not exonic, disruption. Tolerance could be related to insertions simply being spliced out, as it is known that *Y. lipolytica* splicing can remove relatively large introns compared to other fungi (Mekouar, Blanc-Lenfle et al. 2010). Alternatively, because fungal introns are at the extreme 5' ends of genes, it is possible that effects of intronic insertions are alleviated by secondary downstream transcription or translation start sites. These interpretations are consistent with studies of DNA transposition of the related hAT family transposon piggyBac in *D. melanogaster* that showed that transposition into introns can modify, but not necessarily eliminate gene expression (Hacker, Nystedt et al. 2003). Intronic hits were excluded from the GCS Poisson analysis by subtracting them from H_{gene}, however they were included when calculating the genome-wide rate of gene hits (λ). Second, some genes were represented solely by hits in the extreme downstream end. In previous integration profiling, including Tn3 disruption profiling of *S. cerevisiae* (Ross-Macdonald, Coelho et al. 1999) and Tf1 profiling of *S. pombe* (Guo, Park et al. 2013), 3'-terminal hits were associated with expression of truncated proteins that retained significant activity. In order to avoid false negative findings based on lower phenotypic penetrance of 3' hits, insertions restricted to the downstream 10% of the ORF were subtracted from H_{gene} when they accounted for ≥ 50% of the total H_{gene}. Using the above GCS, genes were classified as essential, growth impaired, non-essential or low-confidence essential (LC essential).   LC-essential genes were those that were sufficiently short so that the probability of 0 or 1 hits was greater than the threshold of $P < 0.05$ and therefore could not be reliably classified as essential. Finally, longer genes occasionally demonstrated an expected vs. observed hit margin sufficient to produce a $P < 0.05$, despite ample hits. Gen0 genes with hit density greater than 3 hits/kb/million reads (HKM) but with $P < 0.05$ were classified as non-essential.

Gen0, Gen20 or Gen80 samples grown in and passaged in either YPD or YPG medium were analyzed as above except inter- and intra–genic hit frequencies (λ) were recalculated for each Gen80 population. After 80 generations, genes could remain non-essential, be reclassified as growth impaired (essential for survival in growth in glucose and glycerol), gly-conditional (essential for survival in glycerol only) or Glu-conditional (essential for survival in glucose only). After the initial Poisson evaluation, logical refining of gene classifications was performed as described above for Gen0 to ensure accuracy. For example, low-flanking hit densities could allow a gene with 1 hit to have a P value > 0.05 and cause the gene to score non-essential. Refining takes into account additional parameters such as gene hit density (HKM) and the overall median of gene hits for essential and non-essential genes to flag possible miscalls. In some cases the density of gene hits at Gen20 are also considered.

Gene disruption

Y. lipolytica genes (*GUT2*, *KU70, MAE3, UTR1*) were disrupted by replacement of the ORF by homologous recombination (HR) using a LoxP-*URA3*-LoxP cassette flanked by sequences 5' and 3' of the ORF. Mutagenized transformants were colony purified and mutants were verified by PCR amplification and DNA sequencing. Disruption of genes (*ADE2*, *GPD1*, *HIS3*, *PEX10*, *SNF1*) was by directed mutagenesis according to the previously described CRISPR method for *Y. lipolytica* (Schwartz, Hussain et al. 2016). For CRISPR, primer sequences were incorporated into *Avr*II-restricted pCRISPRyL by Gibson assembly (Gibson 2009) and plasmids used are listed (Table S1). Predicted essential genes *SNF1* and *HIS3* were disrupted using CRISPR-Cas9 as described above, except in the presence of a *URA3*-marked rescue plasmid containing a wild-type copy of the gene harboring a silent mutation that eliminated the CRISPR guide PAM site.

Gene ontology analysis

GO information was obtained using the Blast2GO software (www.blast2go.com). Using the CLIB89 reference genome (Magnan, Yu et al. 2016), enriched GO terms were determined using the Fisher Exact test ($P < 0.05$) and false discovery rate (FDR) correction. For comparison to *Y. lipolytica*, *S. cerevisiae* essential genes (phenotype:inviable) were collected from the *Saccharomyces* Genome Database (SGD) (Cherry 2016) and *S. pombe* essential genes (FYPO:0002061) were taken from Pombase (2016). GO enrichment for *S. cerevisiae* (sacCer3 reference genome) and *S.*

pombe (ASM294v2 reference genome) was also done using Blast2GO. The Venn diagrams were prepared using venn.js software (https://github.com/benfred/venn.js).

Z-score calculations for mutant representation

Reads per insertion mutation ("hit") were used as a proxy for the representation of mutants in cultures. In this analysis hits with fewer than 10 reads, within introns or in the last 10% of the gene, unless that was the sole hit, were filtered out (raw reads are reported Table S2). Reads per hit in Gen0, Gen20Glu, Gen80Glu, Gen20Gly and Gen80Gly samples were determined, normalized by dividing by total reads per that culture, log2 transformed, and used to determine the average normalized read count and standard deviation for each culture. These were used to determine Z-scores for each hit and these were averaged for the hits in each gene to give a final Z-score average per gene. These values are reported in Table S3 as relative fitness.

Fluorescence-activated cell sorting

A frozen aliquot of Gen0 library cells was thawed, pelleted, resuspended at an OD_{600} =~ 0.1 in SD -Leu (nitrogen-enriched) and grown in flasks with shaking at 200 rpm at 23^{o}C overnight. Cells were inoculated to OD_{600}=0.05 and grown to a final OD=~3.5. For isolation of a lipid-enriched population by cell sorting, cells were stained with the lipophilic fluorescent dye Bodipy [BODIPY® 493/503 (4,4-Difluoro-1,3,5,7,8-Pentamethyl-4-Bora-3a,4a-Diaza-*s*-Indacene, Molecular Probes]. Bodipy was dissolved in DMSO at a concentration of 5mM and stored at -80 ^{o}C until use., One OD_{600} of cells was stained in 1 x PBS containing 8μM Bodipy for 15 min in the dark at room temperature. This procedure was scaled up as needed. Stained cells were washed and resuspended in nitrogen-enriched medium at a concentration of 1 x 10^{6}/ml. Samples were kept on ice and in the dark. Cells were sorted using a JSAN cell sorter (Bay Biosciences, Japan) at a rate ~200 events/sec. Bodipy was excited with a 488 nm laser, and fluorescent emission was captured using a 530/30 nm bandpass filter. Cells were fractionated into SD -Leu medium based on intensity of Bodipy staining (1% highly stained and 99% remaining). Non-stained cells were used as autofluorescence control. Collected cells from both the 1% and 99% stained-cell populations were plated onto YPD. For an initial screen, individual colonies from both fractions were analyzed by Bodipy fluorescence for lipid content. Cells were grown in 96-well format plates to OD_{600}=~4-6. Cells were fixed by addition of 1/9 volume of formaldehyde, incubated at room temperature with shaking for 20 min, washed with 1 x PBS and stored at 4^{o}C until analyzed. Cells were stained with Bodipy as described above and fluorescence intensity was determined in 96-well plate format by fluorescence-activated cell sorting using a NovoCyte flow cytometer (Acea Biosciences, Inc.). Fluorescence data were analyzed using FlowJo V10 software and expressed as the normalized geometric mean fluorescence (MFI). Representative highly-stained strains were grown in nitrogen-enriched medium in flasks with shaking at 200 rpm at 23^{o}C for 48 hrs (exponential growth). Strains were fixed and analyzed in 96-well plate format for Bodipy fluorescence. The data are reported as the average and standard deviation of the geometric mean fluorescence of three individual biological replicates.

Sequencing of insertion site

Transposon insertion locations in mutants identified by Bodipy staining were determined from genomic DNA by inverse PCR (Sambrook and Russell 2006). GDNAs were digested with *Bcu*1, *Nhe*1, *Xba*1 and *Xma*J1 or *Bam*HI, *Bgl*II and *Bcl*1 to generate compatible cohesive ends. Digested gDNAs were circularized by ligation and the transposon-genome junction sequence was amplified by PCR (Table S1). The PCR fragments were cloned into PCR4-Blunt TOPO vector (Invitrogen), the transposon junction sequence determined (Genewiz) using primer VB5510, and mapped to the YALI1 (CLIB89) genome (Magnan, Yu et al. 2016).

Lipid determination

Lipid content of individual mutants isolated from the 1% lipid-rich fraction by cell sorting was determined and compared to the lipid content of the control strain (yJY1948). To maximize lipid production, cell cultures were inoculated to an OD_{600} of 0.5 in 50 ml nitrogen-depleted medium in flasks and grown with shaking at 200 rpm at 28ºC for 120 hrs. A minimum of 4-5 biological replicates were analyzed for each strain. Cell pellets (~25 OD_{600} units) were collected, centrifuged and the supernatant removed and stored at -80 ºC. A duplicate sample was collected, desiccated to complete dryness and weighed to determine the cell dry weight. Fatty acids were extracted from ~25 OD_{600} frozen cell pellets (4-5 biological replicates) to which was added 5 µl of internal standard (nonadecanoic acid; stock solution 0.015 g/ml) in the presence of 500 µl 10% barium hydroxide and 550 µl of 1,4-dioxane at 110ºC for 18-24 hrs. Samples were acidified with 6M HCl to pH<4. Fatty acids were recovered in hexane and concentrated under a stream of N_2 to near dryness. Samples were derivatized by addition of 1 ml of 1N HCl in methanol and heated at 80ºC for 30 min. Samples were cooled, and 1 ml 0.9% NaCl and 1 ml hexane was added, and vortexed for 5 min. Samples were centrifuged at 2,000 g for 10 min at 25ºC, the hexane layer removed and evaporated under N_2 stream to a final volume of ~1 ml. Analysis of derivatized fatty acids was performed by gas chromatography-flame ionization detector (GC-FID) on an Agilent GC 7890A (Agilent Technologies, Santa Clara, CA) equipped with a HP5MSI column (length 30 m, 0.250 mm ID, 0.25 µm film thickness). The oven temperature started at 80ºC, then 20ºC/min to 210ºC, hold 1 min, and 10ºC/min to 280ºC hold 1 min, 20ºC/min up to 320ºC hold for 3 min. The detector–FID was at 250ºC, H_2 flow 30 ml/min, Air 400 ml/min and helium as makeup gas at 25 ml/min. Peaks were identified using F.A.M.E. C8-C24 mix (Sigma-Aldrich).

ROC analysis of genome scale models

Two genome-scale models iNL895 (Loira, Dulermo et al. 2012) and iYali4 (Kerkhoven, Pomraning et al. 2016) were obtained from BioModels Database (2016) as MODEL11111900000 and MODEL MODEL1508190002 respectively. The models were loaded in Cobra Toolbox (Becker, Feist et al. 2007) for Matlab, which was used to iteratively perform flux balance analysis (FBA) simulation on each model so that changes in predicted biomass production after a single gene deletion could be recorded and analyzed downstream. When predicted biomass productions are lower than a set threshold (50% of max biomass production), the corresponding deleted genes are suggested to be essential for fitness, while non-essential otherwise. The biomass predictions were then compared with the genes from Hermes transposon mutagenesis essential/non-essential classification (899 genes in iNL895 and 901 genes in iYali4). For ROC analysis Hermes-derived gene classifications were used as classifiers while biomass (arbitrary units) were used as predictors. Area under the curve (AUC) values with 99% confidence intervals were used to assess performance of each model. ROC analysis was done in R using the pROC package (Robin, Turck et al. 2011).

Materials and Methods described in Chapter 5:

Strains and Plasmids

Yarrowia lipolytica strains are derived from W29/CLIB89. Strain yJL2006 has deletions of both *URA3* and *LEU2* (Patterson et al. 2017); yJY2039 contains integrated 2-PS expression cassette mobilized by Hermes transposition using pJY4089; yJL2008 has a disruption of *PEX10* by CRISPR method using pJL4074. pDTN3997 is an episomal expression plasmid containing three copies of the 2-PS expression cassette. Details of strain and plasmid construction are described in Supporting Information.

Culture conditions

For transformation, cells were cultured in synthetic dextrose medium (SD) [0.67% yeast nitrogen base, 2% dextrose], containing complete amino acids, inositol and adenine sulfate lacking selection nutrients for selection and growth of cells transformed with plasmids containing particular prototrophic markers. Nitrogen-limited medium (NL) was composed of 1.6 g/L yeast nitrogen base without amino acids or ammonium sulfate (Difco), 0.625 g/L ammonium sulfate (Fisher), and 5 g/L sodium sulfate (Fisher). Glycerol (40 g/L ((Fisher) or dextrose (20g/L or 40g/L as noted) (Sigma) was provided as the carbon source. Nitrogen–replete medium (NR) was NL medium supplemented with 5g/L ammonium sulfate. Cerulenin (Sigma-Aldrich) was diluted in DMSO (Sigma-Aldrich) to a 50 mM stock solution and supplemented to medium as noted. Cells were cultures in flasks or bioreactor at 28ºC for the times indicated. Cells were grown in flasks or in a bioreactor as described (Supplemental Information).

TAL analysis

TAL was quantified by reading the absorbance at 282 nm (A_{282}) using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific). A dilution series of pure TAL (Sigma Aldrich) dissolved in minimal medium was used to relate A_{282} to titer. This method was verified using HPLC (Cardenas and Da Silva).

Chemical conversion of TAL into pogostone

In a one-pot procedure, pogostone was synthesized in a reaction containing TAL and 4-methylpentanoic acid in the presence of N,N'-Dicyclohexylcarbodiimide (DCC) and 4-dimethylaminopyridine (DMAP). The reaction, shown in Fig. 5.2a, was carried out under argon at room temperature for 3 hrs, followed by heating at 100ºC for 5 hrs. Purifying the product by flash column chromatography afforded pogostone in 96% yield. The reaction was first carried out in 2.0 mM scale, but later was scaled up to 10 g and gave similar results. Biobased TAL purified from biomass broth (94-100% purity) was used in the same protocol. Analogs of pogostone were produced in the same reaction using the relevant acids and heating at 100ºC to overnight (Fig. 5.2b).

CHAPTER 6: Summary and Conclusions

Developing new technologies to accelerate the bioengineering of *Y. lipolytica*

At the onset of this work, one of the most striking deficiencies missing from the molecular toolkit of *Y. lipolytica* was the lack of a complete genome assembly. To address this, we proposed to build an independent genome assembly of *Y. lipolytica* using traditional short read Illumina technology coupled with contemporary PacBio long read and BioNano genome mapping technologies. This combinatorial approach not only allowed us to assemble the genome into chromosome length contigs but also allowed us to accurately place repetitive sequences and identify intriguing genomic features such as a 54 kb block repeat and several overlapping reading frames.

Subsequent comparisons of the CLIB89 genome with the previous CLIB122 genome assembly identified several striking differences between the genomes of these two closely related strains of *Y. lipolytica*. The Ylt1 and Tyl6 families of Ty3/Gypsy elements are completely absent from CLIB89. Genomic locations of full length Ylli retroelements differ between the two strains, suggesting that this element may still be actively transposing. These findings highlight the importance of accurate genome assemblies for industrial strains of *Y. lipolytica*; assumptions and observations made from the genome assembly of one strain may not be applicable to the other strains. For instance, integration strategies that capitalized on zeta repetitive elements (Ylt1 solo LTRs) would not function in strains lacking the Ylt1 retroelement (Schmid-Berger, Schmid et al. 1994; Fickers, Fudalej et al. 2005).

After assembly of a high quality genome, we sought to identify the essential genes in this organism. To accomplish this goal, we performed saturation mutagenesis with the Hermes DNA cut-and-paste transposable element. Sequencing of all the Hermes insertion sites revealed notable gaps in the genome which we surmised to be essential genes. By mapping these genes onto metabolic pathways, we were able to also assess the necessity of entire pathways. For instance, we discovered that the entire histidine biosynthetic pathway was essential. This pathway also includes *HIS3*, one of the most commonly used nutritional markers in *S. cerevisiae*. Findings such as these may help gain insight into the nutritional requirements necessary for growth in *Y. lipolytica* which may have consequences for future bioengineering of this organism.

A list of essential genes would not only inform and direct future metabolic engineering efforts but would could also be used to assess the validity of genome scale metabolic models. Several studies advocate for the use of *in silico* predictive models, especially for metabolic enzymes that may have a multi-tiered system of transcriptional, translational, and post translational control of expression (Loira, Dulermo et al. 2012; Kerkhoven, Pomraning et al. 2016; Kerkhoven, Kim et al. 2017). Using our essential gene list, we demonstrate that the currently existing metabolic models underestimate the number of essential genes to a similar extent. The development of future metabolic models would be made more accurate by incorporating the essential gene data set identified by our transposon mediated functional genomics approach.

In the course of our work, we also developed several protocols to stably integrate heterologous expression cassettes into the *Y. lipolytica* genome. Unlike *S. cerevisiae* which possesses a very high rate of homologous recombination, *Y. lipolytica* has a demonstrably low rate of targeted integration, even with very long 1 kb flanking sequence. Although exogenous DNA is efficiently incorporated into the genome, integration via non homologous end joining results in random integration sites, which may or may not be favorable for high levels of transcriptional expression. In response, we developed a transposition based integration strategy utilizing the Hermes DNA cut-and-paste transposon. The Hermes transposon favors integration into regions with low nucleosome occupancy, which are regions typically associated with active transcription and represents a promising alternative to random integration. Another alternative integration method would be to use CRISPR to target repetitive elements in the genome. We demonstrate the feasibility of this strategy by utilizing a CRISPR construct that targeted the repetitive Ylli retroelement. This strategy has the potential to quickly insert multiple copies of a gene of interest, multimeric enzyme complexes, or even entire multi enzyme pathways into the *Y. lipolytica* genome. However, in order to assess the efficacy of these alternative integration strategies, GFP can be placed under the control of a *Y. lipolytica* promoter and integrated into the genome by one of these methods. Fluorescence intensity could be quantified with cytometry and, coupled with real time PCR to quantify copy number, the level of expression per copy number for each of these integration strategies could be compared.

Bioengineering of *Y. lipolytica* for the production of a novel polyketide

Many groups have utilized Y. lipolytica and other oleaginous organisms for the microbial production of lipids in nitrogen limited medium. It has been speculated that the ATP citrate lyase pathway is activated upon nitrogen depletion, and that the resulting cytosolic acetyl CoA is directed towards *de novo* fatty acid synthesis and lipid production. We have shown that under nitrogen limitation, *Y. lipolytica* also shows increased TAL production. This result suggests that the acetyl CoA produced in this pathway may be redirected into other pathways that also utilize acetyl CoA. Additionally, the requirement for nitrogen limiting conditions precludes the use of complex medium, which has been shown to be detrimental to the downstream conversion to sorbic acid (Schwartz, Johnson et al. 2014). We show that TAL can be purified from the minimal growth medium in a form that is suitable for downstream conversion into pogostone.

Recent studies have reported that there is particularly strong lipid turnover in wildtype *Y. lipolytica*, and stored triglycerides are oxidized and citric acid and mannitol are produced, even in the complete absence of nitrogen (Papanikolaou, Chatzifragkou et al. 2009; Makri, Fakas et al. 2010; Tchakouteu, Kalantzi et al. 2015). These reports are consistent with our finding that acetyl CoA resulting from constitutively active peroxisomal beta oxidation significantly contributes to TAL titers. In response, several groups have engineered strains of *Y. lipolytica* that are incapable of beta oxidation and these engineered strains are capable of accumulating up to 90% lipid (Blazeck, Hill et al. 2014). Taken together, these results demonstrate the potential flexibility of a *Y. lipolytica* microbial factory. With proper engineering, *Y. lipolytica* can be an efficient host organism not only for the production of microbial lipids, but also for polyketides and other acetyl CoA based molecules.

The results of our efforts to engineer the production of TAL suggest that oleaginous organisms may be better defined, not by the accumulation high levels of lipids, but by the high carbon fluxes into acetyl CoA synthesis under nitrogen limiting conditions. In unengineered cells, this acetyl CoA is directed into fatty acid synthesis, but we show that this acetyl CoA can be directed into other heterologous pathways. Finally, oleaginous organisms are typically classified into a single broad stroke mostly centered on the ACL pathway: leverage the unique ACL pathway to manipulate carbon flux into fatty acid synthesis. However, the findings reported in this work identify the contribution of beta oxidation to TAL synthesis. Instead of focusing on just the ACL pathway, the results of this work warrant further research into the unique metabolic subtleties of each oleaginous organism.

REFERENCES

Acker, J., C. Ozanne, et al. (2008). "Dicistronic tRNA-5S rRNA genes in Yarrowia lipolytica: an alternative TFIIIA-independent way for expression of 5S rRNA genes." Nucleic Acids Res **36**(18): 5832-5844.

Alexander, R. P., G. Fang, et al. (2010). "Annotating non-coding regions of the genome." Nat Rev Genet **11**(8): 559-571.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Amberg, D. C., D. J. Burke, et al. (2005). Methods in Yeast Genetics. Cold Spring Harbor, Cold Spring Harbor Laboratory Press.

Andrews, S. J. and J. A. Rothnagel (2014). "Emerging evidence for functional peptides encoded by short open reading frames." Nat Rev Genet **15**(3): 193-204.

Arensburger, P., R. H. Hice, et al. (2011). "Phylogenetic and functional characterization of the hAT transposon superfamily." Genetics **188**(1): 45-57.

Austin, M. B. and J. P. Noel (2003). "The chalcone synthase superfamily of type III polyketide synthases." Nat Prod Rep **20**(1): 79-110.

Ausubel, F. M., R. Brent, et al. (2007). Current Protocols in Molecular Biology, John Wiley and Sons, Inc.

Bankar, A., A. Kumar, et al. (2009). "Environmental and industrial applications of Yarrowia lipolytica." Appl Microbiol Biotechnol **84**(5): 847-865.

Barth, G. and C. Gaillardin (1996). *Yarrowia lipolytica*. Non-conventional yeasts in biotechnology. A handbook. K. Wolf. Heidelberg, Springer**:** 313-388.

Barth, G. and C. Gaillardin (1997). "Physiology and genetics of the dimorphic fungus Yarrowia lipolytica." FEMS Microbiology Reviews **19**(4): 219-237.

Barth, G., Gaillardin, C. (1996). Yarrowia lipolytica. In: Nonconventional Yeasts in Biotechnology: A Handbook (Wolf, K., Ed.), pp. 313-388. Springer-Verlag.

Baudin, A., O. Ozier-Kalogeropoulos, et al. (1993). "A simple and efficient method for direct gene deletion in Saccharomyces cerevisiae." Nucleic Acids Res **21**(14): 3329-3330.

Beauregard, A., M. J. Curcio, et al. (2008). "The take and give between retrotransposable elements and their hosts." Annu Rev Genet **42**: 587-617.

Becker, S. A., A. M. Feist, et al. (2007). "Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox." Nat Protoc **2**(3): 727-738.

Beopoulos, A., J. Cescut, et al. (2009). "Yarrowia lipolytica as a model for bio-oil production." Prog Lipid Res **48**(6): 375-387.

Beopoulos, A., T. Chardot, et al. (2009). "Yarrowia lipolytica: A model and a tool to understand the mechanisms implicated in lipid accumulation." Biochimie **91**(6): 692-696.

Beopoulos, A., Z. Mrozova, et al. (2008). "Control of lipid accumulation in the yeast Yarrowia lipolytica." Appl Environ Microbiol **74**(24): 7779-7789.

Beopoulos, A., J. M. Nicaud, et al. (2011). "An overview of lipid metabolism in yeasts and its impact on biotechnological processes." Appl Microbiol Biotechnol **90**(4): 1193-1206.

Beopoulos, A., J. Verbeke, et al. (2014). "Metabolic engineering for ricinoleic acid production in the oleaginous yeast Yarrowia lipolytica." Appl Microbiol Biotechnol **98**(1): 251-262.

Blazeck, J., A. Hill, et al. (2014). "Harnessing Yarrowia lipolytica lipogenesis to create a platform for lipid and biofuel production." Nat Commun **5**: 3131.

Boeke, J. D., J. Trueheart, et al. (1987). "5-Fluoroorotic acid as a selective agent in yeast molecular genetics." Methods Enzymol **154**: 164-175.

Boulton, C. A., and Colin Ratledge (1981). "Correlation of lipid accumulation in yeasts with possession of ATP: citrate lyase." J Gen Microbiol **127**(1): 169-176.

Burge, C. B. (1998). Modeling dependencies in pre-mRNA splicing signals. Computational Methods in Molecular Biology. S. Salzberg, D. Searls and S. Kasif. Amsterdam, Elsevier Science**:** 127-163.

Byrne, K. P. and K. H. Wolfe (2005). "The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species." Genome Res **15**(10): 1456-1461.

Calvey, C. H., L. B. Willis, et al. (2014). "An optimized transformation protocol for Lipomyces starkeyi." Curr Genet **60**(3): 223-230.

Cardenas, J. and N. A. Da Silva "Metabolic engineering of Saccharomyces cerevisiae for the production of triacetic acid lactone." Metab Eng **25**: 194-203.

Cardenas, J. and N. A. Da Silva (2014). "Metabolic engineering of Saccharomyces cerevisiae for the production of triacetic acid lactone." Metabolic Engineering **25**: 194-203.

Casaregola, S. and G. Barth (2013). Transposable Elements and Their Activities in Yarrowia Lipolytica. Munster, Germany, Springer.

Casaregola, S., C. Feynerol, et al. (1997). "Genomic organization of the yeast Yarrowia lipolytica." Chromosoma **106**(6): 380-390.

Casaregola, S., C. Neuveglise, et al. (2002). "Ylli, a non-LTR retrotransposon L1 family in the dimorphic yeast Yarrowia lipolytica." Mol Biol Evol **19**(5): 664-677.

Casaregola, S., C. Neuveglise, et al. (2000). "Genomic exploration of the hemiascomycetous yeasts: 17. Yarrowia lipolytica." FEBS Lett **487**(1): 95-100.

Chalvet, F., C. Grimaldi, et al. (2003). "Hop, an active Mutator-like element in the genome of the fungus Fusarium oxysporum." Mol Biol Evol **20**(8): 1362-1375.

Chen, D. C., J. M. Beckerich, et al. (1997). "One-step transformation of the dimorphic yeast Yarrowia lipolytica." Appl Microbiol Biotechnol **48**(2): 232-235.

Chen, D. C., J. M. Beckerich, et al. (1997). "One-step transformation of the dimorphic yeast Yarrowia lipolytica." Appl Microbiol Biotechnol **48**(2): 232-235.

Cherry, M. (2016). "Saccharomyces Genome Database." from http://www.yeastgenome.org/.

Chia, M., M. A. Haider, et al. (2013). "Mechanistic insights into ring-opening and decarboxylation of 2-pyrones in liquid water and tetrahydrofuran." J Am Chem Soc **135**(15): 5699-5708.

Christen, S. and U. Sauer (2011). "Intracellular characterization of aerobic glucose metabolism in seven yeast species by 13C flux analysis and metabolomics." FEMS Yeast Res **11**(3): 263-272.

Clare, J. J., L. S. Davidow, et al. (1986). "Cloning and characterisation of the ribosomal RNA genes of the dimorphic yeast, Yarrowia lipolytica." Curr Genet **10**(6): 449-452.

Coelho, M. A., Amaral, P.F., Belo, I (2010). "Yarrowia lipolytica: An Industrial Workhorse." Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology: 930-940.

Corthesy-Theulaz, I. E., G. E. Bergonzelli, et al. (1997). "Cloning and characterization of Helicobacter pylori succinyl CoA:acetoacetate CoA-transferase, a novel prokaryotic member of the CoA-transferase family." J Biol Chem **272**(41): 25659-25667.

Craig, N. (2015). A moveable feast: an introduction to mobile DNA. Mobile DNA III. N. Craig, M. Chandler, M. Gellertet al. Washington DC, ASM Press**:** 3-39.

Crook, N., J. Abatemarco, et al. (2016). "In vivo continuous evolution of genes and pathways in yeast." Nat Commun **7**: 13051.

Crooks, G. E., G. Hon, et al. (2004). "WebLogo: a sequence logo generator." Genome Res **14**(6): 1188-1190.

Curcio, M. J., S. Lutz, et al. (2015). "The Ty1 LTR-Retrotransposon of Budding Yeast, Saccharomyces cerevisiae." Microbiol Spectr **3**(2): MDNA3-0053-2014.

Daboussi, M. J. and P. Capy (2003). "Transposable elements in filamentous fungi." Annu Rev Microbiol **57**: 275-299.

Daboussi, M. J., T. Langin, et al. (1992). "Fot1, a new family of fungal transposable elements." Mol Gen Genet **232**(1): 12-16.

Davidow, L., D. Apostolakos, et al. (1985). "Integrative transformation of the yeast Yarrowia lipolytica." Current Genetics **10**(1): 39-48.

Dewannieux, M., C. Esnault, et al. (2003). "LINE-mediated retrotransposition of marked Alu sequences." Nat Genet **35**(1): 41-48.

Dimmer, K. S., S. Fritz, et al. (2002). "Genetic basis of mitochondrial function and morphology in Saccharomyces cerevisiae." Mol Biol Cell **13**(3): 847-853.

Dong, T., E. P. Knoshaug, et al. (2016). "Lipid recovery from wet oleaginous microbial biomass for biofuel production: A critical review." Applied Energy **177**: 879-895.

Dowell, R. D., O. Ryan, et al. (2010). "Genotype to phenotype: a complex problem." Science **328**(5977): 469.

Duarte, N. C., M. J. Herrgard, et al. (2004). "Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model." Genome Res **14**(7): 1298-1309.

Dujon, B., D. Sherman, et al. (2004). "Genome evolution in yeasts." Nature **430**(6995): 35-44.

Dulermo, R., T. Onodera, et al. (2015). "Identification of new genes contributing to the extreme radioresistance of Deinococcus radiodurans using a Tn5-based transposon mutant library." PLoS One **10**(4): e0124358.

Dulermo, T., B. Treton, et al. (2013). "Characterization of the two intracellular lipases of Y. lipolytica encoded by TGL3 and TGL4 genes: new insights into the role of intracellular lipases and lipid body organisation." Biochim Biophys Acta **1831**(9): 1486-1495.

Durrens, P. and D. J. Sherman (2005). "A systematic nomenclature of chromosomal elements for hemiascomycete yeasts." Yeast **22**(5): 337-342.

Duvick, J., D. S. Standage, et al. (2016). "xGDBvm: A Web GUI-driven workflow for annotating eukaryotic genomes in the cloud." Plant Cell.

Evertts, A. G., C. Plymire, et al. (2007). "The hermes transposon of Musca domestica is an efficient tool for the mutagenesis of Schizosaccharomyces pombe." Genetics **177**(4): 2519-2523.

Fickers, P., F. Fudalej, et al. (2005). "Selection of new over-producing derivatives for the improvement of extracellular lipase production by the non-conventional yeast Yarrowia lipolytica." Journal of Biotechnology **115**(4): 379-386.

Fickers, P., M. T. Le Dall, et al. (2003). "New disruption cassettes for rapid gene disruption and marker rescue in the yeast Yarrowia lipolytica." J Microbiol Methods **55**(3): 727-737.

Forbes, E. M., S. R. Nieduszynska, et al. (2007). "Control of gag-pol gene expression in the Candida albicans retrotransposon Tca2." BMC Mol Biol **8**: 94.

Fournier, P., L. Guyaneux, et al. (1991). "Scarcity of ars sequences isolated in a morphogenesis mutant of the yeast Yarrowia lipolytica." Yeast **7**(1): 25-36.

Gabus, C., D. Ficheux, et al. (1998). "The yeast Ty3 retrotransposon contains a 5'-3' bipartite primer-binding site and encodes nucleocapsid protein NCp9 functionally homologous to HIV-1 NCp7." EMBO J **17**(16): 4873-4880.

Gaillardin, C., M. Mekouar, et al. (2013). Comparative Genomics of Yarrowia lipolytica. Berlin, Springer-Verlag.

Gaillardin, C., C. Neuveglise, et al. (2012). "Mitochondrial genomes of yeasts of the Yarrowia clade." <u>FEMS Yeast Res</u> **12**(3): 317-331.

Gangadharan, S., L. Mularoni, et al. (2010). "DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo." <u>Proc Natl Acad Sci U S A</u> **107**(51): 21966-21972.

Gasmi, N., A. Ayed, et al. (2011). "Design of an efficient medium for heterologous protein production in Yarrowia lipolytica: case of human interferon alpha 2b." <u>Microbial Cell Factories</u> **10**(1): 38.

Gerdes, S. Y., M. D. Scholle, et al. (2003). "Experimental determination and system level analysis of essential genes in Escherichia coli MG1655." <u>J Bacteriol</u> **185**(19): 5673-5684.

Giaever, G., A. M. Chu, et al. (2002). "Functional profiling of the Saccharomyces cerevisiae genome." <u>Nature</u> **418**(6896): 387-391.

Giaever, G. and C. Nislow (2014). "The yeast deletion collection: a decade of functional genomics." <u>Genetics</u> **197**(2): 451-465.

Gibson, D. G. (2009). "Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides." <u>Nucleic Acids Res</u> **37**(20): 6984-6990.

Gietz, R. D. and R. H. Schiestl (1991). "Applications of high efficiency lithium acetate transformation of intact yeast cells using single-stranded nucleic acids as carrier." <u>Yeast</u> **7**(3): 253-263.

Gietz, R. D. and R. H. Schiestl (2007). "High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method." <u>Nat Protoc</u> **2**(1): 31-34.

Gietz, R. D. and R. H. Schiestl (2007). "Large-scale high-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method." Nat Protoc **2**(1): 38-41.

Goodwin, T. J., J. E. Ormandy, et al. (2001). "L1-like non-LTR retrotransposons in the yeast Candida albicans." Curr Genet **39**(2): 83-91.

Groenewald, M., T. Boekhout, et al. (2014). "Yarrowia lipolytica: safety assessment of an oleaginous yeast with a great industrial potential." Crit Rev Microbiol **40**(3): 187-206.

Guimond, N., D. K. Bideshi, et al. (2003). "Patterns of Hermes transposition in Drosophila melanogaster." Mol Genet Genomics **268**(6): 779-790.

Guo, Y. and H. L. Levin (2010). "High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in Schizosaccharomyces pombe." Genome Res **20**(2): 239-248.

Guo, Y., J. M. Park, et al. (2013). "Integration profiling of gene function with dense maps of transposon integration." Genetics **195**(2): 599-609.

Guthrie, C. and G. R. Fink (1991). Guide to Yeast Genetics and Molecular Biology, Academic Press.

Guthrie, C. and G. R. Fink (1991). Guide to Yeast Genetics and Molecular Biology. San Diego, CA, Academic Press Inc.

Hacker, U., S. Nystedt, et al. (2003). "piggyBac-based insertional mutagenesis in the presence of stably integrated P elements in Drosophila." Proc Natl Acad Sci U S A **100**(13): 7720-7725.

Hansen, L. J., D. L. Chalker, et al. (1988). "Ty3, a yeast retrotransposon associated with tRNA genes, has homology to animal retroviruses." Mol Cell Biol **8**(12): 5245-5256.

Hicks, J. B., A. Hinnen, et al. (1979). "Properties of Yeast Transformation." Cold Spring Harbor Symposia on Quantitative Biology **43**: 1305-1313.

Holton, N. J., T. J. Goodwin, et al. (2001). "An active retrotransposon in Candida albicans." Nucleic Acids Res **29**(19): 4014-4024.

Hong, J. J., T. Y. Wu, et al. (2013). "Viral IRES prediction system - a web server for prediction of the IRES secondary structure in silico." PLoS One **8**(11): e79288.

Hook, M. and X. Tang (2013). "Depletion of fossil fuels and anthropogenic climate change-A review." Energy Policy **52**: 797-809.

Horton, R. M., H. D. Hunt, et al. (1989). "Engineering hybrid genes without the use of restriction enzymes: gene splicing by overlap extension." Gene **77**(1): 61-68.

Hynes, M. J., S. L. Murray, et al. (2008). "Genetic analysis of the role of peroxisomes in the utilization of acetate and fatty acids in Aspergillus nidulans." Genetics **178**(3): 1355-1369.

Kamisaka, Y., K. Kimura, et al. (2013). "Overexpression of the active diacylglycerol acyltransferase variant transforms Saccharomyces cerevisiae into an oleaginous yeast." Appl Microbiol Biotechnol **97**(16): 7345-7355.

Kanehisa, M. and S. Goto (2000). "KEGG: kyoto encyclopedia of genes and genomes." Nucleic Acids Res **28**(1): 27-30.

Kastenmayer, J. P., L. Ni, et al. (2006). "Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae." Genome Res **16**(3): 365-373.

Kerkhoven, E. J., Y. M. Kim, et al. (2017). "Leucine Biosynthesis Is Involved in Regulating High Lipid Accumulation in Yarrowia lipolytica." MBio **8**(3).

Kerkhoven, E. J., Y. M. Kim, et al. (2017). "Leucine Biosynthesis Is Involved in Regulating High Lipid Accumulation in Yarrowia lipolytica." Mbio **8**(3).

Kerkhoven, E. J., K. R. Pomraning, et al. (2016). "Regulation of amino-acid metabolism controls flux to lipid accumulation in *Yarrowia lipolytica*." Systems Biology and Applications **2**.

Kerkhoven, E. J., K. R. Pomraning, et al. (2016). "Regulation of amino-acid metabolism controls flux to lipid accumulation in Yarrowia lipolytica." NPJ Syst Biol Appl **2**: 16005.

Kerscher, S., G. Durstewitz, et al. (2001). "The complete mitochondrial genome of yarrowia lipolytica." Comp Funct Genomics **2**(2): 80-90.

Kim, D. U., J. Hayles, et al. (2010). "Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe." Nat Biotechnol **28**(6): 617-623.

Kovalchuk, A., S. Senam, et al. (2005). "Tyl6, a novel Ty3/gypsy-like retrotransposon in the genome of the dimorphic fungus Yarrowia lipolytica." Yeast **22**(12): 979-991.

Kramerov, D. A. and N. S. Vassetzky (2011). "Origin and evolution of SINEs in eukaryotic genomes." Heredity (Edinb) **107**(6): 487-495.

Kraus, G. A., U. K. Wanninayake, et al. (2016). "Triacetic acid lactone as a common intermediate for the synthesis of 4-hydroxy-2-pyridones and 4-amino-2-pyrones." Tetrahedron Letters **57**(11): 1293-1295.

Kretschmer, M., J. Wang, et al. (2012). "Peroxisomal and Mitochondrial beta-Oxidation Pathways Influence the Virulence of the Pathogenic Fungus Cryptococcus neoformans." Eukaryotic Cell **11**(8): 1042-1054.

Kretzschmar, A., C. Otto, et al. (2013). "Increased homologous integration frequency in Yarrowia lipolytica strains defective in non-homologous end-joining." Curr Genet **59**(1-2): 63-72.

Kretzschmar, A., C. Otto, et al. (2013). "Increased homologous integration frequency in Yarrowia lipolytica strains defective in non-homologous end-joining." Current Genetics **59**(1-2): 63-72.

Kumar, A. and M. Snyder (2001). "Genome-wide transposon mutagenesis in yeast." Curr Protoc Mol Biol **Chapter 13**: Unit13 13.

Kunau, W. H., V. Dommes, et al. (1995). "beta-oxidation of fatty acids in mitochondria, peroxisomes, and bacteria: a century of continued progress." Prog Lipid Res **34**(4): 267-342.

Kurian, J. V. (2005). "A new polymer platform for the future - Sorona (R) from corn derived 1,3-propanediol." Journal of Polymers and the Environment **13**(2): 159-167.

Lang, F. "RNAweasel and MFannot: preddiction of complex sturctured mitochondrial RNAs, using ERPIN (1) as a search engine.", from http://megasun.bch.umontreal.ca/RNAweasel/.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.

Larkin, M. A., G. Blackshields, et al. (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

Li, M. A., S. J. Pettitt, et al. (2013). "The piggyBac transposon displays local and distant reintegration preferences and can cause mutations at noncanonical integration sites." Mol Cell Biol **33**(7): 1317-1330.

Li, Q., W. Du, et al. (2008). "Perspectives of microbial oils for biodiesel production." Appl Microbiol Biotechnol **80**(5): 749-756.

Li, Y. H., Z. B. Zhao, et al. (2007). "High-density cultivation of oleaginous yeast Rhodosporidium toruloides Y4 in fed-batch culture." Enzyme and Microbial Technology **41**(3): 312-317.

Li, Z., H. Sun, et al. (2013). "Overexpression of malic enzyme (ME) of Mucor circinelloides improved lipid accumulation in engineered Rhodotorula glutinis." Appl Microbiol Biotechnol **97**(11): 4927-4936.

Lisch, D. (2002). "Mutator transposons." Trends Plant Sci **7**(11): 498-504.

Liu, L. and H. S. Alper (2014). "Draft Genome Sequence of the Oleaginous Yeast Yarrowia lipolytica PO1f, a Commonly Used Metabolic Engineering Host." Genome Announc **2**(4).

Liu, L., A. Pan, et al. (2015). "An evolutionary metabolic engineering approach for enhancing lipogenesis in Yarrowia lipolytica." Metab Eng **29**: 36-45.

Liu, X. Y., Z. Chi, et al. (2013). "Both Decrease in ACL1 Gene Expression and Increase in ICL1 Gene Expression in Marine-Derived Yeast Yarrowia lipolytica Expressing INU1 Gene Enhance Citric Acid Production from Inulin." Marine Biotechnology **15**(1): 26-36.

Loira, N., T. Dulermo, et al. (2012). "A genome-scale metabolic model of the lipid-accumulating yeast Yarrowia lipolytica." BMC Syst Biol **6**(1): 35.

Loira, N., T. Dulermo, et al. (2012). "A genome-scale metabolic model of the lipid-accumulating yeast Yarrowia lipolytica." BMC Syst Biol **6**: 35.

Lomsadze, A., V. Ter-Hovhannisyan, et al. (2005). "Gene identification in novel eukaryotic genomes by self-training algorithm." Nucleic Acids Res **33**(20): 6494-6506.

Lopes, M. L., S. C. Paulillo, et al. (2016). "Ethanol production in Brazil: a bridge between science and industry." Braz J Microbiol **47 Suppl 1**: 64-76.

Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic Acids Res **25**(5): 955-964.

Madzak, C., J.-M. Nicaud, et al. (2005). Yarrowia lipolytica. Production of Recombinant Proteins, Wiley-VCH Verlag GmbH & Co. KGaA**:** 163-189.

Maggio-Hall, L. A. and N. P. Keller (2004). "Mitochondrial beta-oxidation in Aspergillus nidulans." Mol Microbiol **54**(5): 1173-1185.

Magnan, C., J. Yu, et al. (2016). "Sequence Assembly of Yarrowia lipolytica Strain W29/CLIB89 Shows Transposable Element Diversity." PLoS One **11**(9): e0162363.

Makri, A., S. Fakas, et al. (2010). "Metabolic activities of biotechnological interest in Yarrowia lipolytica grown on glycerol in repeated batch cultures." Bioresour Technol **101**(7): 2351-2358.

Malik, H. S. and T. H. Eickbush (1999). "Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons." J Virol **73**(6): 5186-5190.

Mansour, S. L., K. R. Thomas, et al. (1988). "Disruption of the proto-oncogene int-2 in mouse embryo-derived stem cells: a general strategy for targeting mutations to non-selectable genes." Nature **336**(6197): 348-352.

Marck, C., R. Kachouri-Lafond, et al. (2006). "The RNA polymerase III-dependent family of genes in hemiascomycetes: comparative RNomics, decoding strategies, transcription and evolutionary implications." Nucleic Acids Res **34**(6): 1816-1835.

Mekouar, M., I. Blanc-Lenfle, et al. (2010). "Detection and analysis of alternative splicing in Yarrowia lipolytica reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts." Genome Biol **11**(6): R65.

Merz, S. and B. Westermann (2009). "Genome-wide deletion mutant analysis reveals genes required for respiratory growth, mitochondrial genome maintenance and mitochondrial protein synthesis in Saccharomyces cerevisiae." Genome Biol **10**(9): R95.

Molin, M., J. Norbeck, et al. (2003). "Dihydroxyacetone kinases in Saccharomyces cerevisiae are involved in detoxification of dihydroxyacetone." J Biol Chem **278**(3): 1415-1423.

Morin, N., J. Cescut, et al. (2011). "Transcriptomic analyses during the transition from biomass production to lipid accumulation in the oleaginous yeast Yarrowia lipolytica." PLoS One **6**(11): e27966.

Muller, S., T. Sandal, et al. (1998). "Comparison of expression systems in the yeasts Saccharomyces cerevisiae, Hansenula polymorpha, Klyveromyces lactis, Schizosaccharomyces pombe and Yarrowia lipolytica. Cloning of two novel promoters from Yarrowia lipolytica." Yeast **14**(14): 1267-1283.

Nakamura, C. E. and G. M. Whited (2003). "Metabolic engineering for the microbial production of 1,3-propanediol." Curr Opin Biotechnol **14**(5): 454-459.

Neuveglise, C., F. Chalvet, et al. (2005). "Mutator-like element in the yeast Yarrowia lipolytica displays multiple alternative splicings." Eukaryot Cell **4**(3): 615-624.

Neuveglise, C., H. Feldmann, et al. (2002). "Genomic evolution of the long terminal repeat retrotransposons in hemiascomycetous yeasts." Genome Res **12**(6): 930-943.

Neuveglise, C., C. Gaillardin, et al. (2013). Noncoding RNA genes transcribed by RNA Pol III in Yarrowia lipolytica genome.

Nicaud, J. M. (2012). "Yarrowia lipolytica." Yeast **29**(10): 409-418.

Nicaud, J. M., E. Fabre, et al. (1989). "Expression of invertase activity in Yarrowia lipolytica and its use as a selective marker." Curr Genet **16**(4): 253-260.

Nicaud, J. M., C. Madzak, et al. (2002). "Protein expression and secretion in the yeast Yarrowia lipolytica." FEMS Yeast Res **2**(3): 371-379.

Nissen, T. L., U. Schulze, et al. (1997). "Flux distributions in anaerobic, glucose-limited continuous cultures of Saccharomyces cerevisiae." Microbiology **143 (Pt 1)**: 203-218.

Nookaew, I., M. C. Jewett, et al. (2008). "The genome-scale metabolic model iIN800 of Saccharomyces cerevisiae and its validation: a scaffold to query lipid metabolism." BMC Syst Biol **2**: 71.

Ogrydziak, D. M. and S. J. Scharf (1982). "Alkaline Extracellular Protease Produced by Saccharomycopsis lipolytica CX161–1B." Journal of General Microbiology **128**(6): 1225-1234.

Oldenburg, K. R., K. T. Vo, et al. (1997). "Recombination-mediated PCR-directed plasmid construction in vivo in yeast." Nucleic Acids Res **25**(2): 451-452.

Omura, S. (1976). "The antibiotic cerulenin, a novel tool for biochemistry as an inhibitor of fatty acid synthesis." Bacteriol Rev **40**(3): 681-697.

Papanikolaou, S., A. Chatzifragkou, et al. (2009). "Biosynthesis of lipids and organic acids by Yarrowia lipolytica strains cultivated on glucose." European Journal of Lipid Science and Technology **111**(12): 1221-1232.

Pfeiffer, T. and A. Morley (2014). "An evolutionary perspective on the Crabtree effect." Front Mol Biosci **1**: 17.

Pomraning, K. R. and S. E. Baker (2015). "Draft Genome Sequence of the Dimorphic Yeast Yarrowia lipolytica Strain W29." Genome Announc **3**(6).

Pomraning, K. R., E. L. Bredeweg, et al. (2017). "Regulation of Nitrogen Metabolism by GATA Zinc Finger Transcription Factors in Yarrowia lipolytica." Msphere **2**(1).

Pomraning, K. R., S. Wei, et al. (2015). "Comprehensive Metabolomic, Lipidomic and Microscopic Profiling of Yarrowia lipolytica during Lipid Accumulation Identifies Targets for Increased Lipogenesis." PLoS One **10**(4): e0123188.

Proux-Wera, E., D. Armisen, et al. (2012). "A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach." BMC Bioinformatics **13**: 237.

Przybyla-Zawislak, B., R. A. Dennis, et al. (1998). "Genes of succinyl-CoA ligase from Saccharomyces cerevisiae." Eur J Biochem **258**(2): 736-743.

Qian, W., D. Ma, et al. (2012). "The genomic landscape and evolutionary resolution of antagonistic pleiotropy in yeast." Cell Rep **2**(5): 1399-1410.

Qiao, K., S. H. Imam Abidi, et al. (2015). "Engineering lipid overproduction in the oleaginous yeast Yarrowia lipolytica." Metab Eng **29**: 56-65.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Ratledge, C. (2002). "Regulation of lipid accumulation in oleaginous micro-organisms." Biochem Soc Trans **30**(Pt 6): 1047-1050.

Ratledge, C. and J. P. Wynn (2002). "The biochemistry and molecular biology of lipid accumulation in oleaginous microorganisms." Adv Appl Microbiol **51**: 1-51.

Reid, I., N. O'Toole, et al. (2014). "SnowyOwl: accurate prediction of fungal genes by using RNA-Seq and homology information to select among ab initio models." BMC Bioinformatics **15**: 229.

Richardson, S. R., A. J. Doucet, et al. (2015). "The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes." Microbiol Spectr **3**(2).

Robertson, D. S. (1978). "Characterization of the Mutator System in Maize." Mutat. Res **51**: 21-28.

Robin, X., N. Turck, et al. (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves." BMC Bioinformatics **12**: 77.

Ross-Macdonald, P., P. S. Coelho, et al. (1999). "Large-scale analysis of the yeast genome by transposon tagging and gene disruption." Nature **402**(6760): 413-418.

Runguphan, W. and J. D. Keasling (2014). "Metabolic engineering of Saccharomyces cerevisiae for production of fatty acid-derived biofuels and chemicals." Metab Eng **21**: 103-113.

Sambrook, J. and D. W. Russell (2006). Molecular Cloning, Cold Spring Harbor Laboratory Press.

Sandmeyer, S., K. Patterson, et al. (2015). "Ty3, a Position-specific Retrotransposon in Budding Yeast." Microbiol Spectr **3**(2).

Saunders, L. P., M. J. Bowman, et al. (2015). "Triacetic acid lactone production in industrial Saccharomyces yeast strains." Journal of Industrial Microbiology & Biotechnology **42**(5): 711-721.

Schapira, M., M. Tyers, et al. (2017). "WD40 repeat domain proteins: a novel target class?" Nat Rev Drug Discov **16**(11): 773-786.

Schmid-Berger, N., B. Schmid, et al. (1994). "Ylt1, a highly repetitive retrotransposon in the genome of the dimorphic fungus Yarrowia lipolytica." J Bacteriol **176**(9): 2477-2482.

Schneider, T. D. and R. M. Stephens (1990). "Sequence logos: a new way to display consensus sequences." Nucleic Acids Res **18**(20): 6097-6100.

Schwartz, C. M., M. S. Hussain, et al. (2016). "Synthetic RNA Polymerase III Promoters Facilitate High-Efficiency CRISPR-Cas9-Mediated Genome Editing in Yarrowia lipolytica." ACS Synth Biol.

Schwartz, T. J., R. L. Johnson, et al. (2014). "Engineering Catalyst Microenvironments for Metal-Catalyzed Hydrogenation of Biologically Derived Platform Chemicals." Angewandte Chemie-International Edition **53**(47): 12718-12722.

Seip, J., R. Jackson, et al. (2013). "Snf1 is a regulator of lipid accumulation in Yarrowia lipolytica." Appl Environ Microbiol **79**(23): 7360-7370.

Sestric, R., G. Munch, et al. (2014). "Growth and neutral lipid synthesis by Yarrowia lipolytica on various carbon substrates under nutrient-sufficient and nutrient-limited conditions." Bioresour Technol **164**: 41-46.

Shafiee, S. and E. Topal (2009). "When will fossil fuel reserves be diminished?" Energy Policy **37**(1): 181-189.

Sherman, D., P. Durrens, et al. (2004). "Genolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts." Nucleic Acids Res **32**(Database issue): D315-318.

Sherman, D. J., T. Martin, et al. (2009). "Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes." Nucleic Acids Res **37**(suppl 1): D550-D554.

Sherman, F. (2002). "Getting started with yeast." Methods Enzymol. **350B**: 3-41.

Slater, G. S. and E. Birney (2005). "Automated generation of heuristics for biological sequence comparison." BMC Bioinformatics **6**: 31.

Souciet, J., M. Aigle, et al. (2000). "Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies." FEBS Lett **487**(1): 3-12.

Sprague, G. F. and J. E. Cronan (1977). "Isolation and characterization of Saccharomyces cerevisiae mutants defective in glycerol catabolism." J Bacteriol **129**(3): 1335-1342.

Standage, D. S. (2010-2015). "AEGeAn:an integrated toolkit for analysis and evaluation of annotated genomes." from http://standage.github.io/AEGeAn.

Stanke, M. and B. Morgenstern (2005). "AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints." Nucleic Acids Res **33**(Web Server issue): W465-467.

Steinmetz, L. M., H. Sinha, et al. (2002). "Dissecting the architecture of a quantitative trait locus in yeast." Nature **416**(6878): 326-330.

Sun, J. and H. S. Alper (2015). "Metabolic engineering of strains: from industrial-scale to lab-scale chemical production." J Ind Microbiol Biotechnol **42**(3): 423-436.

Swamy, M. K. and U. R. Sinniah (2015). "A Comprehensive Review on the Phytochemical Constituents and Pharmacological Activities of Pogostemon cablin Benth.: An Aromatic Medicinal Plant of Industrial Importance." Molecules **20**(5): 8521-8547.

Symer, D. E. and J. D. Boeke (2010). An everlasting war dance between retrotransposons and their metazoan hosts. Retroviruses. R. Kurth and N. Bannert. Norfolk, UK, Caister Academic Press**:** 1-34.

Tai, M. and G. Stephanopoulos (2013). "Engineering the push and pull of lipid biosynthesis in oleaginous yeast Yarrowia lipolytica for biofuel production." Metab Eng **15**: 1-9.

Tang, X. L., H. X. Feng, et al. (2013). "Metabolic engineering for enhanced fatty acids synthesis in Saccharomyces cerevisiae." Metabolic Engineering **16**: 95-102.

Tchakouteu, S. S., O. Kalantzi, et al. (2015). "Lipid production by yeasts growing on biodiesel-derived crude glycerol: strain selection and impact of substrate concentration on the fermentation efficiency." Journal of Applied Microbiology **118**(4): 911-927.

Teng, X., M. Dayhoff-Brannigan, et al. (2013). "Genome-wide consequences of deleting any single gene." Mol Cell **52**(4): 485-494.

Ter-Hovhannisyan, V., A. Lomsadze, et al. (2008). "Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training." Genome Res **18**(12): 1979-1990.

Tomaszewska, L., A. Rywinska, et al. (2012). "Production of erythritol and mannitol by Yarrowia lipolytica yeast in media containing glycerol." Journal of Industrial Microbiology & Biotechnology **39**(9): 1333-1343.

Trapnell, C., L. Pachter, et al. (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.

UniProt, C. (2015). "UniProt: a hub for protein information." Nucleic Acids Res **43**(Database issue): D204-212.

van Heerikhuizen, H., A. Ykema, et al. (1985). "Heterogeneity in the ribosomal RNA genes of the yeast Yarrowia lipolytica; cloning and analysis of two size classes of repeats." Gene **39**(2-3): 213-222.

van Opijnen, T., K. L. Bodi, et al. (2009). "Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms." Nat Methods **6**(10): 767-772.

Vassetzky, N. S. and D. A. Kramerov (2013). "SINEBase: a database and tool for SINE analysis." Nucleic Acids Res **41**(Database issue): D83-89.

Verbeke, J., A. Beopoulos, et al. (2013). "Efficient homologous recombination with short length flanking fragments in Ku70 deficient Yarrowia lipolytica strains." Biotechnology Letters **35**(4): 571-576.

Walsh, D. and I. Mohr (2011). "Viral subversion of the host protein synthesis machinery." Nat Rev Microbiol **9**(12): 860-875.

Wang, C., Z. Wang, et al. (2017). "The augmented lipid productivity in an emerging oleaginous model alga Coccomyxa subellipsoidea by nitrogen manipulation strategy." World J Microbiol Biotechnol **33**(8): 160.

Wang, G., D. Li, et al. (2018). "Comparative transcriptome analysis reveals multiple functions for Mhy1p in lipid biosynthesis in the oleaginous yeast Yarrowia lipolytica." Biochim Biophys Acta **1863**(1): 81-90.

Wang, J.-H., W. Hung, et al. (2011). "High efficiency transformation by electroporation of Yarrowia lipolytica." The Journal of Microbiology **49**(3): 469-472.

Wang, Z. P., H. M. Xu, et al. (2013). "Disruption of the MIG1 gene enhances lipid biosynthesis in the oleaginous yeast Yarrowia lipolytica ACA-DC 50109." Biochim Biophys Acta **1831**(4): 675-682.

Weber, H. and C. Kurischko (1989). "Sexual behaviour in the alkane-utilizing yeast Yarrowia lipolytica." Yeast **5 Spec No**: S279-285.

Winzeler, E. A., D. D. Shoemaker, et al. (1999). "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis." Science **285**(5429): 901-906.

Wiser, M. J. and R. E. Lenski (2015). "A Comparison of Methods to Measure Fitness in Escherichia coli." PLoS One **10**(5): e0126210.

Wolf, K. (1996). Nonconventional Yeasts in Biotechnology. Berlin, Germany, Springer Berlin Heidelberg.

Wolf, K. (1996). Nonconventional yeasts in biotechnology: a handbook. Berlin, Bermany, Springer-Verlag.

Wood, V., R. Gwilliam, et al. (2002). "The genome sequence of Schizosaccharomyces pombe." Nature **415**(6874): 871-880.

Xu, P., K. Qiao, et al. (2017). "Engineering oxidative stress defense pathways to build a robust lipid production platform in Yarrowia lipolytica." Biotechnol Bioeng **114**(7): 1521-1530.

Xuan, J.-W., P. Fournier, et al. (1988). "Cloning of the LYS5 gene encoding saccharopine dehydrogenase from the yeast Yarrowia lipolytica by target integration." Current Genetics **14**(1): 15-21.

Yadav, V. G. and G. Stephanopoulos (2014). "Metabolic engineering: the ultimate paradigm for continuous pharmaceutical manufacturing." ChemSusChem **7**(7): 1847-1853.

Yanisch-Perron, C., J. Vieira, et al. (1985). "Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors." Gene **33**(1): 103-119.

Yim, H., R. Haselbeck, et al. (2011). "Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol." Nat Chem Biol **7**(7): 445-452.

Yoon, O. K. and R. B. Brem (2010). "Noncanonical transcript forms in yeast and their regulation during environmental stress." RNA **16**(6): 1256-1267.

Yu, S. F., D. N. Baldwin, et al. (1996). "Human foamy virus replication: a pathway distinct from that of retroviruses and hepadnaviruses." Science **271**(5255): 1579-1582.

Yuzbasheva, E. Y., E. B. Mostova, et al. (2017). "A metabolic engineering strategy for producing free fatty acids by the Yarrowia lipolytica yeast based on impairment of glycerol metabolism." Biotechnol Bioeng.

Zeng, W., H. Zhang, et al. (2017). "Biosynthesis of keto acids by fed-batch culture of Yarrowia lipolytica WSH-Z06." Bioresour Technol **243**: 1037-1043.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Res **18**(5): 821-829.

Zheng, L., U. Baumann, et al. (2004). "An efficient one-step site-directed and site-saturation mutagenesis protocol." Nucleic Acids Res **32**(14): e115.

Zhou, Y. J., N. A. Buijs, et al. (2014). "Fatty Acid-Derived Biofuels and Chemicals Production in Saccharomyces cerevisiae." Front Bioeng Biotechnol **2**: 32.

Zhu, Q. and E. N. Jackson (2015). "Metabolic engineering of Yarrowia lipolytica for industrial applications." Curr Opin Biotechnol **36**: 65-72.