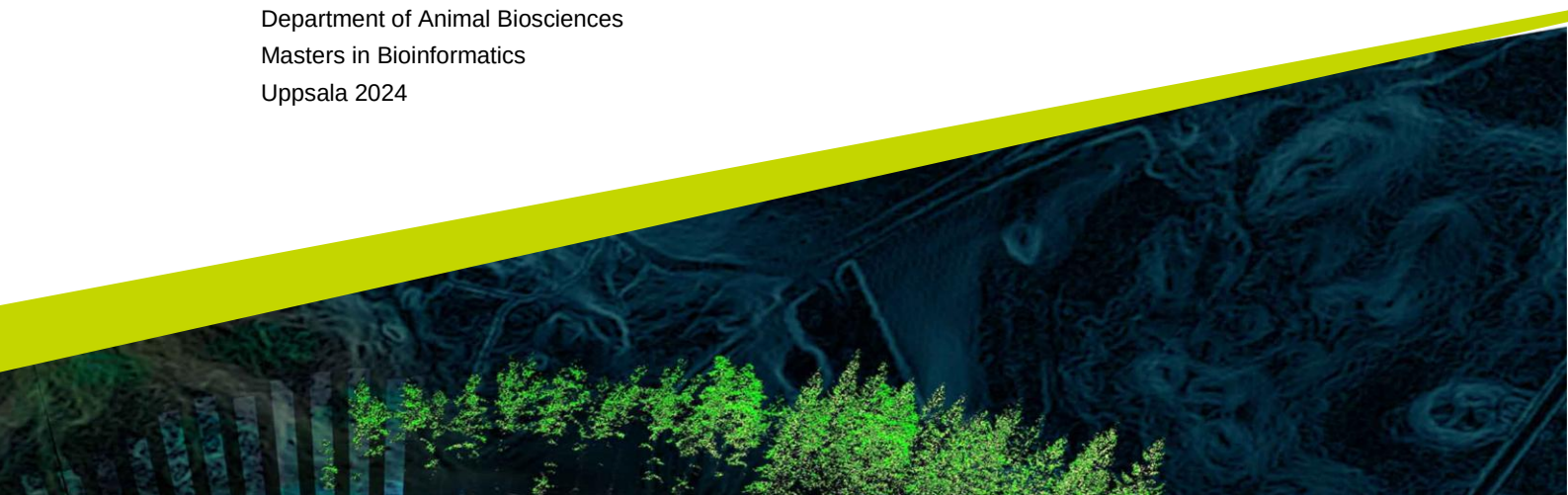
Genetic and Metabolic Analysis of *Rhodotorula toruloides* Strains for Enhanced Lipid Production

Sven Kristian Feddersen

Independent project in Bioinformatics • 30 credits
Swedish University of Agricultural Sciences, SLU
Department of Animal Biosciences
Masters in Bioinformatics
Uppsala 2024



Genetic and Metabolic Analysis of *Rhodotorula toruloides* Strains for Enhanced Lipid Production

Sven Kristian Feddersen

Supervisor:	**Bettina Müller, SLU, Department of Molecular Sciences**
Examiner:	Samuel Coulbourn Flores, SLU, Department of Animal Breeding and Genetics, Bioinformatics

Credits:	30 credits
Level:	Second-cycle, A2E
Course title:	Independent Project in Bioinformatics
Course code:	EX1002
Programme/education:	Masters in Bioinformatics
Course coordinating dept:	Department of Animal Biosciences
Place of publication:	Uppsala
Year of publication:	2024

| **Keywords:** | Oleaginous, yeast, Rhodotorula, toruloides, genome, alignment |

Swedish University of Agricultural Sciences
Department of Animal Biosciences, Bioinformatics

Abstract

In the context of sustainable biofuel production, oleaginous yeasts are emerging as critical players in the search for profitable, greener alternatives to fossil fuels. These microbes, known for their powerful lipid synthesis capabilities, have the potential to convert renewable, low-value biomass into economically valuable lipids. This process offers potential profit by significantly lowering production costs compared to conventional lipid sources, such as vegetable oils, and mitigating associated environmental impacts. Among these, the Rhodotorula species stand out for their strong lipid production skills and genetic diversity (Osman et al., 2022)(Zhang et al., 2021). The need to identify replacements for traditional lipid sources cannot be overstressed. Traditional methods are not only unsustainable but also have significant environmental consequences. Oleaginous yeasts provide a possible alternative since they convert low-value, non-food biomass into high-value lipids. This skill presents them as a sustainable alternative for waste management and the production of biofuels, chemicals, and food additives (Zhang et al., 2021). Rhodotorula yeasts, in particular, are known for their ability to use a wide range of substrates, including waste products and raw plant materials, which has the potential to drastically transform the existing economic landscape toward sustainability. This master's thesis dives into the genetic complexities of three Rhodotorula strains—two parental strains (CBS 14 and CBS 349) and one hybrid strain (CBS 6016), with the goal of investigating the genetic and metabolic capabilities of the hybrid strain CBS 6016, focusing on the inheritance and functionality of key metabolic pathways, particularly those involved in lipid production. Our findings show that CBS 6016 inherits a large number of protein sequences from both parental strains, preserving essential metabolic functions, as well as a partial loss of some enzymes suggests potential areas where metabolic capabilities could be affecting overall fitness and growth.

Keywords: Oleaginous, yeast, *Rhodotorula, toruloides,* peptide, alignment.

Table of contents

List of figures

Abbreviations

CBS	Centraalbureau voor Schimmelcultures
DBVPG	Industrial Yeasts Collection DBVPG
NGR	A specific strain of oleaginous red yeast
LDP1	Lipid droplet protein 1 gene
ACC1	Acetyl Coa Carboxylase gene
FAS1	Fatty Acid Synthase β subunit gene
DGA1	Diacylglycerol Acyltransferase 1
TAG	Triacylglyceride
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
Cas9	CRISPR associated protein 9
IFO	Institute for Fermentation, Osaka
HH	Hemicellulosic Hydrolysate
CG	Crude Glycerol
ONT	Oxford Nanopore Technologies
FastA	Fast-All
BLAST	Basic Local Alignment Search Tool
KEGG	Kyoto Encyclopedia of Genes and Genomes
ORF	Open Reading Frame
COV	Enzyme Commission
CoA	Coenzyme A
TCA cycle	Tricarboxilic Acid Cycle (Citric Acid Cycle)

1. Literature Review

1.1 Oleaginous Yeasts in Sustainable Applications

Oleaginous yeasts are microorganisms that can accumulate high amounts of lipids. Production can reach 20% of their dry mass of cells and can get to 70% under suitable conditions. The traditional sources of lipids contribute significantly to the environmental burden, leading to the destruction of forests and the generation of greenhouse gases. In light of this, oleaginous yeast are a viable and sustainable alternative. They can effectively convert a variety of substrates, especially lignocellulosic materials and crude glycerol (a by-product of the biodiesel industry) to such valuable lipids whose applications are as diverse as biofuels, oleochemicals, and food additives, indicating the potential of oleaginous yeasts in contributing to a more sustainable future (Adrio, 2017; Abeln & Chuck, 2021; Passoth et al., 2023).

Oleaginous yeasts are not only engaged in producing lipids but also because the ability to utilize low-cost, non-food resources of renewable nature such as lignocellulosic biomass ensures sustainability of oleaginous yeasts and ensures environmental challenges' resilience to improve cost-effectiveness through the conversion of waste into value. This would go a long way in solving environmental problems associated with waste disposal, as well as in developing a sustainable production cycle in which the use of resources is optimized. The development of genetic engineering and metabolic pathway optimization has significantly facilitated the increase in yield and productivity of lipids. All novel tailored yeasts for the applications allow for more cost- and energy-effective processes in producing lipids. Today, researchers can adapt oleaginous yeasts to the applications they would like and easily revolutionize the biotechnological landscape (Adrio, 2017; Abeln & Chuck, 2021).

This has thus underlined the biotechnological potential of oleaginous yeasts since it could entail new and innovative ways to reduce dependence on non-renewable resources and to promote the circular bioeconomy. In fact, a lot of research and development is being done in this area, with great promises for finding new strains and metabolic engineering strategies. These efforts are providing the stage for novel applications in biofuels, biochemicals, and the food industry, supporting global sustainability and environmental conservation efforts (Adrio, 2017; Abeln & Chuck, 2021; Passoth et al., 2023).

Oleaginous yeasts form a part and parcel of the research in the development of biotechnological solutions towards a sustainable world. Those areas of research,

therefore, have a lot of significance because of the potential of such yeast to the ability to produce lipids from renewable resources coupled with environmental and economic benefits.

In the context of green biotechnology, the exploration of oleaginous yeasts emerges as a pivotal area of research, offering substantial contributions to the production of microbial oil, vibrant carotenoids, and biofuels. Villegas-Méndez et al. (2022) illuminate the path toward sustainable bioprocesses through the efficient valorization of food waste, employing oleaginous yeasts in an innovative amalgamation of solid-state and submerged fermentations. This strategy exemplifies the essence of the circular bioeconomy, adeptly converting low-cost inputs into valuable biocompounds.

Further elaboration by Abeln et al. underscores the versatile functionality of oleaginous yeasts in biofuel generation, demonstrating their capacity to adeptly transform diverse substrates into lipids. This multifaceted utility highlights the yeasts' significant role in fostering a more sustainable future, through practices aimed at minimizing environmental impact, enhancing the yield and productivity of critical biocompounds, and developing scalable, cost-effective bioprocesses. Analyses have shown the feasibility and environmental merit of these biotechnological approaches. Both Villegas-Méndez et al. (2022) and Abeln et al. advocate for enhanced policy support and the creation of innovative business models to boost the bio-based economy. Such frameworks are crucial for commercializing and popularizing sustainable bioprocesses, positioning oleaginous yeasts as essential agents in connecting the gap between economic affordability and environmental responsibility.

1.2 Ploidy Level

Yeast cells can exist in various ploidy states—haploid, diploid, polyploid, and aneuploid—each affecting the organism's fitness and adaptability. Aneuploidy, in particular, is highlighted as a natural evolutionary mechanism, reflecting the genetic flexibility yeast cells exhibit in response to environmental changes (Crandall et al., 2023)(Gerstein and Otto, 2009). Sexual and non-sexual reproduction in yeast are key for understanding how the different forms change and how these changes affect mix-breed yeasts (Crandall et al., 2023). From these reproductions, yeast can create complex pair structures such as dikaryons (double-nuclei) and heterokaryons (mixed-nuclei). These can lead to aneuploidy hybrids due to genetic instability or mismatch between parent types.

Current gene studies, like short-read sequencing and k-mer frequency assesment, helped explore yeast strains' ploidy. Surprising discoveries like tetraploidy in some strains show how complex yeast genetics is (Weiß et al., 2018). These reveal how crucial ploidy levels are in studying the transfer and change of genes in mixed strains.

1.3 Genomic Insights of the Rhodotorula Species

The genomic insights of Rhodotorula species represent a significant step forward in increasing the biotechnological potential of oleaginous yeasts. The research conducted by Martín-Hernández et al. (2022) on *Rhodotorula babjevae*, showed nearly chromosome-level genome assemblies for two strains, CBS 7808 and DBVPG 8058. This study focuses on the genetic diversity within the species, revealing significant intraspecific divergence that could impact lipid synthesis and accumulation. By combining short-read and long-read sequencing data, the researchers achieved a comprehensive understanding of the genome structure, identifying 7,591 and 7,481 protein-coding genes for strains CBS 7808 and DBVPG 8058, respectively. Such in-depth genomic insights are crucial for developing genetic tools aimed at maximizing lipid production and enhancing the biotechnological applications of oleaginous yeasts.

Furthermore, the analysis revealed tetraploidy and strain-specific extrachromosomal endogenous DNA, suggesting that these genetic complexities might enhance the yeasts' ability to efficiently convert various substrates into valuable lipids. The observed genomic divergence between the two strains indicates the potential for further exploration in the quest for robust strains with improved lipid production efficiency. This genomic exploration directly supports efforts to tailor oleaginous yeasts for specific biotechnological applications, as discussed earlier, by providing a genetic foundation for metabolic engineering and strain improvement strategies that could significantly increase lipid yields and productivity (Martín-Hernández et al., 2022).

Following the insights into Rhodotorula species, a deeper understanding of the genetic diversity among oleaginous yeasts is studied by the genome sequencing and comparative analysis of *Sporobolomyces pararoseus* NGR (Li et al., 2020). This study marks a significant leap in revealing the biotechnological potential harbored within the genetic frameworks of oleaginous yeasts. Unlike the broader characterization provided by Martín-Hernández et al. (2022), the investigation into *S. pararoseus* NGR unravels specific genes related to the synthesis of lipids and carotenoids, offering precise targets for metabolic engineering (Martín-Hernández et al., 2022; Li et al., 2020).

Notably, the research delineates the phylogenetic relationships and evolutionary adaptability of oleaginous yeasts within the Sporidiobolales order. By pinpointing the evolutionary trajectory from Sporobolomyces to Rhodotorula through Rhodosporidiobolus, it enriches our comprehension of yeast adaptability and potential for lipid production enhancements. Such phylogenetic insights complement the previously discussed intraspecific divergence within Rhodotorula, providing a broader evolutionary context for oleaginous yeast diversity (Li et al., 2020).

Through the lens of *S. pararoseus* NGR's genome, this segment enriches the narrative on oleaginous yeasts' potential, emphasizing the importance of genetic diversity in unlocking novel biotechnological applications. It harmoniously extends

the discussion from the general biotechnological potential of oleaginous yeasts, their sustainability, and ploidy variations, to a focused exploration of genetic diversity's role in enhancing lipid production capabilities. As such, the genetic insights from *S. pararoseus* NGR not only improves our understanding of oleaginous yeasts' capabilities but also pave the way for innovative approaches in leveraging their full potential for sustainable biotechnology solutions (Li et al., 2020).

1.4 Metabolic Engineering and Strain Improvement

Metabolic engineering in Rhodotorula strains is showing great promise as a way to boost lipid production, taking advantage of their genetic diversity and how they respond differently to various growth substrates. Intronic sequences, especially those from genes involved in making or storing lipids, are turning out to be key in controlling how genes are expressed, significantly affecting the strength of promoters (Schultz et al., 2022). Using strong intronic promoters, like those from the perilipin/lipid droplet protein 1 gene (LDP1), acetyl-CoA carboxylase gene (ACC1), and fatty acid synthase β subunit gene (FAS1), has proven to increase lipid accumulation, especially under conditions of nitrogen starvation, without relying on the usual pathways for oil biosynthesis and storage (Schultz et al., 2022). Furthermore, ramping up the expression of certain enzymes, such as diacylglycerol acyltransferase 1 (DGA1), with these intronic promoters has been shown to greatly enhance lipid content. This underscores the potential of these promoters for effective metabolic engineering in oleaginous Rhodotorula and similar yeast species (Schultz et al., 2022).

In a similar vein, fine-tuning CRISPR/Cas9 gene editing in *Rhodotorula toruloides* IFO0880 has opened up new avenues for gene overexpression and deletion to boost fatty alcohol production (J. Carl Schultz, Cao and Zhao, 2019). Notably, removing genes involved in triacylglyceride (TAG) formation and increasing the expression of genes that boost the production of cytosolic acetyl-CoA and malonyl-CoA have led to significant increases in fatty alcohol levels. Targeting specific acyltransferases, such as DGA1 and LRO1, has proven particularly effective, highlighting their competitive roles in using acyl-CoA molecules (Schultz et al., 2022). In order to maximize the production of lipids and fatty alcohols in Rhodotorula strains, it is crucial to comprehend metabolic pathways in depth and manipulate essential enzyme activities. This is highlighted by the deliberate combination of genetic modifications.

1.5 Lipid Synthesis from Crude Glycerol and Lignocellulosic Biomass

Microbial lipids are produced from lignocellulosic biomass and crude glycerol, which are waste from the production of biofuel and paper mills. According to recent

research, we can employ these waste products and specific yeasts, such as Rhodotorula, to generate lipids. Making biofuels and other products in this way might be more environmentally friendly. According to a research report, certain species of Rhodotorula, primarily *Rhodotorula glutinis* and *Rhodotorula toruloides*, can produce lipids using hemicellulosic hydrolysate (HH) and raw glycerol (CG). When both are present, these yeasts may grow well on these substrates and produce more lipids. This demonstrates a combined effect on the lipid-producing metabolic pathways. Rhodotorula strains were studied, showing they can use different food sources. They are especially good at using glycerol to boost how much fat they make (Chmielarz et al., 2021). Beyond that, scientists studied how this yeast uses a mix of CG and HH. The combination of sugars found in hemicellulose and glycerol can help increase fat production. The ability to use all the different food sources really showcases their metabolic flexibility. The fats produced by these yeasts are similar to vegetable fats. This is good news for biofuel production, as these fats could potentially be a greener source of energy.

1.6 Methods of Bioinformatics in Yeast Genetics

Bioinformatics is a valuable tool in genomics that goes much beyond data management, it is essential for understanding the complex nature of genetic information. The fields of genomics and computational biology have produced advanced techniques that can be used to understand these complexities such as the genetic structures and inheritance patterns of organisms like oleaginous yeasts. One such bioinformatics tool is Diamon, a system created for large-scale genomic sequence alignment quickly. It makes it easier to perform the comparative analysis required to determine which genetic segments are conserved and which are unique among strains that are closely related (Marçais et al., 2018). Diamond's ability to precisely navigate the large genomic landscape allows researchers to identify regions of high homology and synteny, which sheds light on the evolutionary processes that mold yeast strains' genetic composition (Marçais et al., 2018). In particular, in the study of hybrids such as the *Rhodotorula toruloides* species complex, it is important to determine whether parts of the hybrid genome have been inherited from its parental lineages. This analysis plays a critical role in illuminating the history of genomic inheritance. Extensive genomic alignments provide important insights into gene transfer, recombination processes, and ploidy variations—all of which are essential to comprehending the metabolic and adaptable traits of these yeasts.

The genetic foundations of characteristics like lipid production can be examined in detail using bioinformatics as a lens. The capacity to analyze and comprehend the genomic basis of oleaginous yeasts is becoming more and more crucial as scientists work to improve the biotechnological uses of these organisms (Pang et al., 2019)(Jiang et al., 2022). By providing specific methods for metabolic

engineering and strain enhancement, the data derived from these alignments are pushing oleaginous yeasts to the forefront of sustainable biotechnology research.

A complete model of yeast metabolism was built by combining functional genomics, metabolic profiling, and sequence alignment data (Jiang et al., 2022) to better understand how lipid production is controlled. This approach helped pinpoint key control points and pathways that affect lipid yields, essentially creating a roadmap for more effective metabolic engineering. The model can help guide future work on developing sustainable, eco-friendly solutions for renewable energy and material production by boosting lipid synthesis.

1.7 Future Directions

Research on oleaginous yeasts creates new opportunities for biotechnological innovation focused on sustainable lipid synthesis, especially when seen through the lense of genomic insights and sophisticated haplotype phasing tools. Research on oleaginous yeast is an area with great potential for addressing important issues in industrial biotechnology, environmental sustainability, and biofuel generation. The present status of research points in several promising areas for the future:

- Exploring Genomic Diversity: Future research ought to concentrate on taking use of this genetic variety to find and create novel strains with increased lipid production and stress tolerance (Sitepu et al., 2014) as we continue to understand the genetic foundations of lipid synthesis in oleaginous yeasts. In this effort, targeted genome editing techniques like CRISPR-Cas9 will be essential because they allow for precise alterations that may unleash these microbes' unrealized biotechnological potential.
- Improving Metabolic Pathway: The combination of haplotype phasing and metabolic engineering provides a robust framework for optimizing lipid biosynthesis pathways (Zhu et al., 2012). Future research efforts aim to construct comprehensive metabolic models. These models, combined with phasing information, are pivotal for predicting and increasing lipid production, offering a strategic path to increased efficiency.
- Sustainable Feedstock Utilization: The capacity of oleaginous yeasts to convert waste materials into useful lipids is a symbol of the shift towards a circular bioeconomy. The investigation of yeasts that can take in a larger variety of sustainable feedstocks will be given priority in future research. This strategy opens the door to a more environmentally friendly future as well as promoting economic and environmental sustainability (Ratledge, 2014).

By going in these new areas, scientists will be able to fully utilize oleaginous yeasts' potential and advance the creation of biotechnologically sustainable solutions to urgent global issues.

2. Introduction

In today's world, where sustainability is important, oleaginous yeasts are becoming very important in finding greener solutions. These microorganisms, which are good at making lipids, have the ability to turn renewable resources into valuable lipids. This can help reduce our reliance on fossil fuels and lessen the environmental impact of traditional lipid production methods. Among these yeasts, the Rhodotorula species are especially notable for their strong ability to produce lipids and their genetic variety (Osman et al., 2022)(Zhang et al., 2021). The importance of finding alternatives to traditional lipid sources cannot be overstated. Traditional methods are not only unsustainable but also have serious environmental effects. Oleaginous yeasts offer a possible alternative because they can convert low-value, non-food biomass into high-value lipids. This makes them a sustainable option for waste management, and for the production of biofuels, chemicals, and food additives (Zhang et al., 2021). Rhodotorula yeasts, in particular, are known for their ability to use many different types of materials, including waste and raw plant materials, which could greatly change the current economic situation towards sustainability.

The overall aim of this master's thesis is to investigate genetic diversities in three of the Rhodotorula strains; CBS 14, CBS 349, and CBS 6016 that is the hybrid strain derived from the two parental strains. In this study, the major objective is to investigate on the genetic and metabolic potential of CBS 6016 hybrid strain with reference to the genes and their function in the identified metabolic pathways especially in lipid synthesis. In a way, this work's goal is to establish which of the genes the hybrid acquires from its parental strains and which it does not, as it seeks to investigate the processes that may enable the hybrid to synthesise lipids and assess its suitability for biotechnological purposes.

3. Materials and Methods

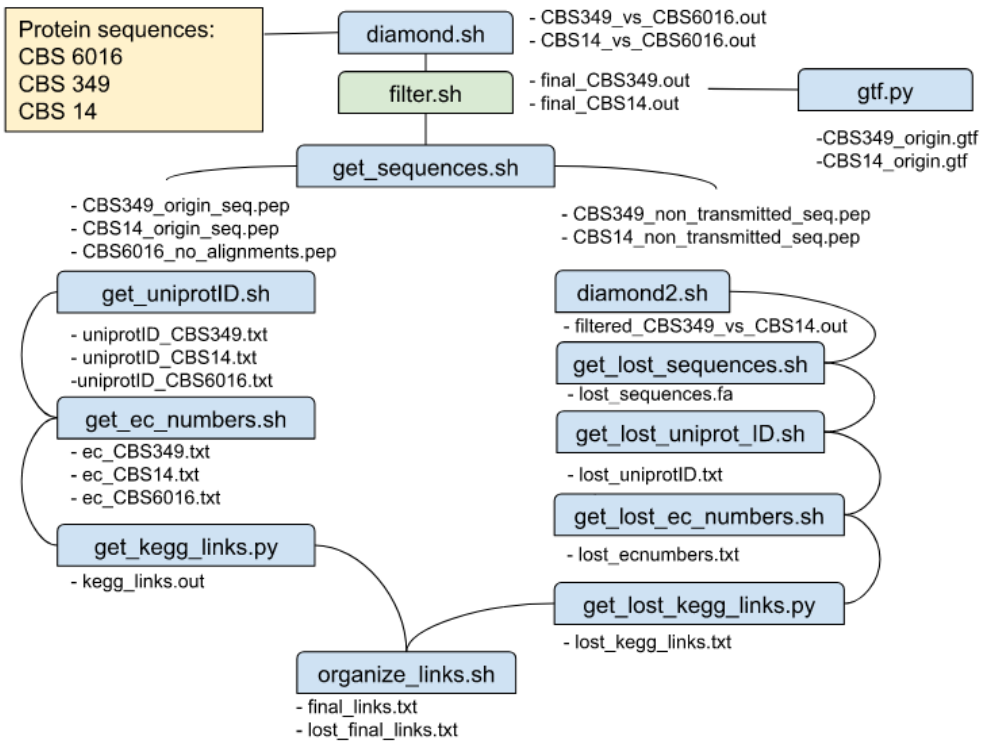


Figure 1. Workflow of the Methods employed in the present study. A diagram illustrating the step-by-step workflow employed in the study.

3.1 Dataset

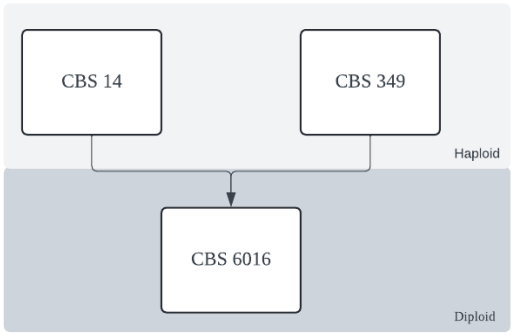


Figure 2. Relationship between the different R. toruloides strains. A diagram showing the relationship between the three R. toruloides strains used in the study. It depicts the parental strains CBS 14 and CBS 349 and their hybrid CBS 6016. Arrows indicate the hybridization process and genetic lineage.

The whole peptide sequences of three *Rhodotorula toruloides* strains—CBS 14, CBS 349, and their hybrid, CBS 6016—make up the dataset used in the present study. While CBS 6016 is a diploid strain produced by the hybridization of the two parental strains, CBS 14 and CBS 349 are both haploid strains. These genetic material used were originally sequenced using Oxford Nanopore Technologies (ONT), which is renowned for producing long reads. This sequencing technology improves the accuracy of structural variation detection and makes it easier to assemble complex genomic areas.

The data originated from a genomic investigation carried out by Martin-Hernandez in 2023 with the goal of investigating the genetic organization and adaptive evolutionary mechanisms of *Rhodotorula toruloides*.

Additionally, Gene Transfer File (GTF) from CBS 6016 has been used to display the best alignments with each of the parental strains.

All scripts used in the analyses of this study can be found at https://github.com/svenkf/Rhodotorula-toruloides .

3.2 Software and Database

3.2.1 DIAMOND (Version 2.0.14)

DIAMOND (Double Index Alignment of Next-generation sequencing Data) is a high-performance tool designed for ultra-fast protein sequence alignment. It was developed to address the speed limitations of traditional BLASTP, achieving

alignment speeds up to 20,000 times faster while maintaining high sensitivity and accuracy. This speed is particularly beneficial for large-scale proteomic studies and applications involving extensive datasets. It uses a seed-and-extend-based alignment method, which helps quickly pinpoint matching regions between sequences.

In this project, DIAMOND was used to efficiently align protein sequences from different strains. Its rapid processing capabilities enabled comprehensive comparisons and the identification of high-similarity sequences across large datasets. DIAMOND's ability to handle extensive sequence databases and support for GPU acceleration further enhanced its performance. Similar to BLASTP, DIAMOND offers multiple output formats and customizable parameters, allowing for tailored analysis.

3.2.2 Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG is an important database that connects genomic information with knowledge about gene function, which makes it possible to better understand gene activity in an organized manner. KEGG provides relevant insights into biological systems and cellular functions, which are essential for developing translational research and drug discovery initiatives.

The primary purpose of using KEGG in this project was to understand the functional implications of the protein sequences that were identified as either inherited or lost during hybridization events. By mapping these sequences to KEGG pathways, I could infer their roles in various metabolic and regulatory processes, which is crucial for understanding the biological impact of the hybridization events.

3.3 Transmitted sequences

We analyzed three strains of *Rhodotorula toruloides*, with two haploid strains (CBS 14 and CBS 349) and one diploid strain (CBS 6016) resulting from the hybridization of the former two. Our workflow began with the protein sequences in FASTA format from these strains.

The initial step involved performing an alignment with DIAMOND using the script “diamond.sh”. This script performed two alignments: CBS 14 against CBS 6016 and CBS 349 against CBS 6016, with CBS 6016 serving as the reference strain and the other two as query strains. The output from this script consisted of two files:

CBS349_vs_CBS6016.out

CBS14_vs_CBS6016.out

Filtering Alignments

First, we processed the total set of alignments to retain only the best alignments based on a combined score, which was calculated as:

$$Combined\ Score = Similarity\ Percentage\ x\ Alignment\ Length$$

This step ensured that for each reference sequence in CBS 6016, only the alignment with the highest combined score was kept. This approach allowed us to prioritize alignments that were not only highly similar but also covered a significant portion of the query sequence.

Once the best alignments were identified and retained, we further filtered these alignments by applying two additional criteria:

- Similarity percentage of 90% or more.
- Coverage of 70% or more, calculated as:

$$Coverage = \frac{Alignment\ Length}{Total\ length\ of\ query\ sequence} x100$$

This final filtering step ensured that the retained alignments were both highly accurate and represented a substantial portion of the sequence being aligned. The resulting filtered alignments were those that met the highest standards of similarity and coverage, reflecting the most reliable matches between the query strains (CBS 14 and CBS 349) and the reference strain (CBS 6016).

Sequence Retrieval

After obtaining the filtered alignments, the next step was to retrieve the specific protein sequences corresponding to these alignments. This process was performed using a custom shell script that extracted both matched and non-matched sequences from the respective strains.

Matched Sequences:

A custom Bash script (get_sequences.sh) was used for sequence retrieval, employing standard Unix utilities such as *awk*, *grep*, *sed*, and *sort*. The script began by identifying the unique sequence IDs from the filtered alignment files for CBS 349 and CBS 14. These IDs correspond to sequences in the reference strain, CBS

6016, that aligned with either CBS 349 or CBS 14. Using these IDs, the script extracted the matched sequences from the CBS 6016 peptide file. A function named extract_sequences() was implemented within the script to read a list of sequence IDs from a file, use awk to parse the reference FASTA file for headers that match these IDs, and output the corresponding sequences. These extracted sequences were saved in two separate files:

- CBS349_origin_seq.pep
- CBS14_origin_seq.pep.

Non-Aligned Sequences from CBS 6016:

Finally, the script identified sequences in CBS 6016 that did not have any corresponding alignments with either CBS 349 or CBS 14. These sequences, which were not aligned with any sequence from the haploid strains, were extracted and saved in the file:

- CBS6016_no_alignments.pep.

Non-Transmitted Sequences:

The script also identified and extracted sequences from CBS 349 and CBS 14 that did not align with any sequence in CBS 6016. This was accomplished by comparing the sequences in the original peptide files of CBS 349 and CBS 14 against the filtered alignment results. The non-matched sequences were saved as:

- CBS349_non_transmitted_seq.pep
- CBS14_non_transmitted_seq.pep

This sequence retrieval process allowed us to categorize the protein sequences based on their transmission from the parental haploid strains (CBS 349 and CBS 14) to the diploid hybrid strain (CBS 6016). The resulting sequence files provided a clear picture of which sequences were transmitted, which were not, and which sequences in the hybrid strain did not have a counterpart in the parental strains.

UniProt IDs retrieval

After retrieving the protein sequences, we aimed to find the corresponding UniProt IDs for the sequences derived from the three Rhodotorula toruloides strains: CBS 349, CBS 14, and CBS 6016.

Database Preparation:

We first downloaded the UniProt database for Rhodotorula toruloides (taxonomy ID: 5533). This database was downloaded in FASTA format and used to create a DIAMOND database, which served as the reference for subsequent alignment

steps. The database creation and sequence alignment were performed using the DIAMOND tool, which is optimized for high-speed sequence alignment.

Sequence Alignment with DIAMOND:
For each set of sequences—CBS349_origin_seq.pep, CBS14_origin_seq.pep, and CBS6016_no_alignments.pep—we ran a DIAMOND BLASTP search against the UniProt Rhodotorula database. The output of these DIAMOND searches was stored in three separate files:

- uniprotID_CBS349.txt for sequences aligned from CBS 349,
- uniprotID_CBS14.txt for sequences aligned from CBS 14, and
- uniprotID_CBS6016.txt for sequences unique to CBS 6016 that did not align with either parental strain.

This process allowed us to map the sequences from the R. toruloides strains to their corresponding UniProt entries, providing a foundation for subsequent analysis.

Enzyme Commission numbers retrieval

Following the identification of UniProt, the next step was finding the corresponding Enzyme Commission (EC) numbers, which gives a standardized classification of the functions of the proteins.

We used the "get_ec_numbers.sh" script to access the UniProt API (Application Programming Interface) and retrieve the EC numbers (Enzyme Commission numbers) for the proteins that had an assigned EC number.

Generating KEGG URLs

Using the obtained EC numbers, we deployed the "get_kegg_links.py" script to generate URLs linking to KEGG web pages. These pages display the metabolic pathways in which the corresponding EC numbers are involved. To facilitate better visualization and analysis, the script assigned specific colors to EC numbers based on their origin:

EC numbers exclusive to CBS 14 were highlighted in blue.
EC numbers exclusive to CBS 349 were highlighted in red.
EC numbers exclusive to CBS 6016 were highlighted in yellow.

In cases where an EC number was present in both CBS 14 and CBS 349, it was colored in purple, indicating that both parental strains contributed a sequence for that EC number. This color-coding system enabled clear differentiation between

the origins of the enzymes in the metabolic pathways, providing insights into how metabolic functions were inherited or shared between the parental strains.

Extraction and Visualization of Gene Annotations from GTF Files

In this analysis, we aimed to visualize the gene annotations of the Rhodotorula toruloides CBS 6016 genome in relation to its alignments with the haploid parental strains CBS 349 and CBS 14. To accomplish this, we used a Python script to filter and extract relevant gene annotations from a GTF file based on the alignment results obtained from the "filter.sh" script.

The process began by using the filtered alignment files generated by "filter.sh" to identify gene IDs in the CBS 6016 genome that aligned with sequences from CBS 349 and CBS 14. These gene IDs were then fed into the Python script, which filtered the CBS 6016 GTF file to create two new GTF files: CBS349_origin.gtf and CBS14_origin.gtf. These files contain the gene annotations specifically corresponding to sequences in CBS 6016 that aligned with CBS 349 and CBS 14, respectively.

Furthermore, the script identified gene annotations in CBS 6016 that did not align with either parental strain, and these were saved in CBS6016_no_alignments.gtf. The newly generated GTF files were subsequently visualized using IGV, allowing us to compare the gene content transmitted from the parental strains and to identify regions in the CBS 6016 genome that lacked alignment to either CBS 349 or CBS 14.

3.4 Non Transmitted Sequences

To identify the sequences lost during the hybridization of CBS 14 and CBS 349 into the diploid strain CBS 6016, we focused on analyzing the sequences that were not transmitted from either parental strain to CBS 6016. The starting point of this analysis was the two files containing sequences from the parental strains that did not align with CBS 6016:

CBS349_non_transmitted_sequences.pep
CBS14_non_transmitted_sequences.pep

The primary objective was to determine which sequences were absent in both files, indicating that these sequences were not transmitted from either parental strain and, therefore, are missing in CBS 6016. Identifying these non-inherited sequences

allowed us to map the corresponding EC numbers and generate KEGG pathway URLs, providing insight into the metabolic functions that may have been lost during the hybridization event.

Alignment and Filtering

We employed “diamond2.sh”, to align the non-transmitted sequences of CBS 349 against those of CBS 14. This step was crucial to determine which sequences were absent in both strains and therefore lost in CBS 6016. The alignment results were further filtered to ensure they met the similarity and coverage criteria. The sequences identified as non-transmitted by both CBS 14 and CBS 349 were considered lost during the hybridization event. To refine our results, we applied a filtering step to retain only alignments with at least 80% similarity and 70% coverage. The 80% similarity threshold was chosen based on prior studies (Martin-Hernandez, 2023), indicating that CBS 14 and CBS 349 are approximately 80% similar at the nucleotide level, ensuring that only highly similar sequences were considered.

Identification of Lost Sequences

Next, we aimed to extract these specific sequences for further analysis. This extraction was accomplished using the script “get_lost_sequences.sh”. The script's primary function was to retrieve the sequences corresponding to the non-transmitted alignments identified in the previous step. It began by extracting the sequence IDs from the filtered alignment results (filtered_CBS349_vs_CBS14.out), which contain the sequence IDs of the alignments representing genetic material lost in the hybridization process. Using these extracted IDs, the script then searched through a FASTA file (hybrid_flye_pilon_metaeuk_stringtie.fna.transdecoder.pep) containing the complete set of protein sequences from the hybrid strain. For each ID, it identified and extracted the corresponding sequence from the FASTA file. The extracted sequences were then saved in a new FASTA file, lost_sequences.fa, which contains only the sequences that were identified as lost during the hybridization of CBS 14 and CBS 349 into CBS 6016.

UniProt ID Retrieval for Lost Sequences

With the lost_sequences.fa file, we aimed to find the UniProt IDs for these sequences. The “get_lost_uniprot_ID.sh” script performed an alignment search using these lost sequences as the query against the Rhodotorula genus database (the

same database used in “get_uniprot_ID.sh”). The result was the lost_uniprot_ID.txt file.

EC Number Retrieval for Lost Sequences

Using the lost_uniprot_ID.txt file, which contains the UniProt IDs, we employed the “get_lost_ec_numbers.sh” script to retrieve the corresponding EC numbers. The script is meant to access UniProt database via its API, which allows access, through the command line, to a series of information, including functional information, a repository of protein and EC numbers. For each UniProt ID listed in the file, the script searches the UniProt database to collect EC numbers, and writes it to a new column of the input file.

KEGG Links Generation for Lost Sequences

With the obtained EC numbers, the “get_lost_kegg_links.py” script generated URLs linking to KEGG web pages. These pages display the metabolic pathways involving the corresponding EC numbers. For better visualization, EC numbers for the lost sequences were highlighted in green to signify the enzymes missing in CBS 6016.

Final Organization

The organize_links.sh script organized all the URLs based on the KEGG map codes they were present in, producing the final output files final_links.txt and lost_final_links.txt. These files provided a comprehensive view of the metabolic pathways, highlighting the contributions from each parental strain and identifying the enzymes missing in the hybrid strain CBS 6016.

The script specifically selected pathways related to lipid metabolism and essential cellular activities due to the particular interest in the yeast species *Rhodotorula toruloides.* This species is known for its ability to produce and accumulate lipids, making it valuable for biotechnological applications. Therefore, the pathways included in the analysis are crucial for understanding lipid synthesis, energy metabolism, and cell respiration.

By organizing the KEGG pathway URLs into categories such as Lipid Synthesis, Energy Metabolism, and Cell Respiration, the output files offer an easy access URLs related important metabolic processes. This organization helps in identifying which specific metabolic functions have been inherited or lost during the hybridization of CBS 14 and CBS 349 into CBS 6016, providing valuable insights for further research and biotechnological exploitation of *Rhodotorula toruloides*.

4. Results

4.1 Gene Distribution Across Contigs in CBS 6016

To investigate the origin of genes within the hybrid strain CBS 6016, we analysed the alignments of its genome with those of the haploid parental strains CBS 14 and CBS 349. After the alignments were filtered and the corresponding information was extracted from the GTF files, they were visualized using Integrative Genomics Viewer (IGV), where genes from CBS 6016 that aligned more closely with CBS 14 are represented in blue, while those that aligned better with CBS 349 are shown in red.

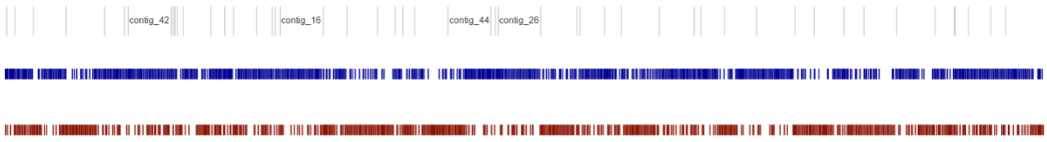


Figure 3. Complete genome of CBS 6016 aligned against its parental strains (CBS 14 in blue, CBS 349 in red). Each horizontal bar represents a contig from the CBS 6016 assembly, with colors indicating alignment to the parental genomes. Blue segments show regions inherited from CBS 14, red segments show regions inherited from CBS 349, and white gaps indicate areas where no alignment was detected. This figure highlights the mosaic nature of the hybrid strain, in which different contigs derive from one or both parents.

We've chosen to focus on contigs 23, 29, and 19 because they each show a different inheritance pattern in CBS 6016: contig 23 mostly aligns with CBS 14 (blue), contig 29 has a more balanced mix of both parents, and contig 19 is again largely from CBS 14 but includes some regions matching CBS 349. Can you explain why the contig numbers are not sequential, in the figure? Why these three contigs?

Contig 23 revealed a distribution with a predominant alignment to CBS 14. The blue track, indicating CBS 14-origin genes, is dense and spans across the entire contig, suggesting a significant contribution from CBS 14 in this region. In contrast, the red track (CBS 349-origin genes) is present but less frequent, indicating a relatively smaller contribution from CBS 349 to this contig.

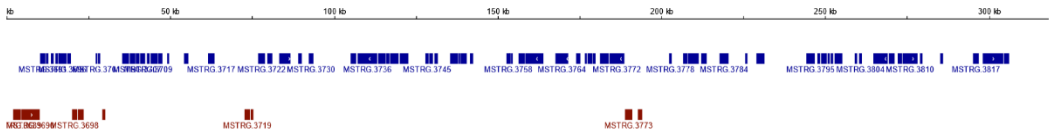


Figure 4. Contig 23 from CBS 6016.

Contig 29 shows a more mixed pattern, with regions of CBS 14-origin genes (blue) and CBS 349-origin genes (red). The presence of both parental contributions in close proximity suggests recombination events during the hybridization process, resulting in a mosaic of genes derived from both parents. Although contig 23 is predominantly aligned with CBS 14, smaller segments also align to CBS 349. These smaller segments of CBS 349-origin within contig 23 suggest localized recombination events. Therefore, while contig 23 mainly reflects inheritance from CBS 14, it is not completely free of recombination. This shows the complexity and mosaic nature of the hybrid genome, where recombination can vary greatly, even within single contigs. But contig 23 also shows recombinations??

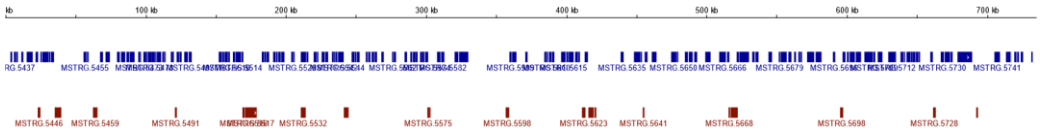


Figure 5. Contig 29 from CBS 6016.

Contig 19 showed a pattern similar to contig 23, with a dominant alignment to CBS 14, indicated by long continuous stretches of blue segments. There are also shorter, scattered red segments aligned to CBS 349. By noting these patterns, we see evidence for limited recombination. This approach allows for a more objective comparison of inheritance patterns across contigs, based on the length, continuity, and distribution of segments inherited from each parent.

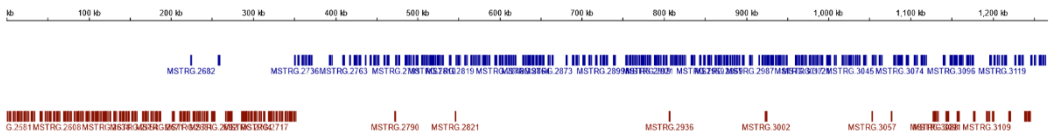


Figure 6. Contig 19 from CBS 6016.

4.2 KEGG Pathways Analysis

The results from the KEGG pathways selected by the *organize_links.sh* script are shown, with a focus on the pathways of most interest to the yeast *Rhodotorula toruloides*. This species is known for the ability to produce and accumulate lipids, making it valuable for biotechnological uses. The pathways chosen in our analysis include fatty acid degradation, oxidative phosphorylation, the citrate cycle (TCA cycle), and glycolysis/gluconeogenesis.

We have used different colors to highlight the origins of a given EC number in the hybrid strain CBS 6016. These colors provide a visual representation of how each enzyme in CBS 6016 is inherited or possibly lost during the hybridization process:

- Purple: Enzymes highlighted in purple indicate that both parental strains, CBS 14 and CBS 349, contributed to the EC number associated with that enzyme.
- Red: Enzymes in red are those that were inherited exclusively from CBS 349.
- Blue: Represent enzymes that were inherited only from CBS 14. Good that the colors are consistent with those of other figures
- Yellow: Enzymes shown in yellow are present in CBS 6016 but did not meet the filtering criteria applied during the analysis. While these enzymes exist in the hybrid genome, their origin (either from one of the parental strains or by recombination, mutation, insertion) is unknown.
- Green: Enzymes highlighted in green indicate potential losses during the hybridization event. These enzymes were not transmitted by either parental strain, suggesting that CBS 6016 may lack these specific metabolic functions.

Fatty Acid Biosynthesis

The ability for the degradation of fatty acids is well developed in *Rhodotorula toruloides*, especially for lipid metabolism. In CBS 6016, several enzymes in this pathway were inherited for a normal function. The KEGG pathway images reveal that while many enzymes required for fatty acid biosynthesis are present in CBS 6016, the inheritance pattern varies. A few enzymes were inherited from both parental strains, and some exclusively from CBS 14. These enzymes, ensure that essential steps in fatty acid synthesis, such as the elongation and termination of fatty acid chains, can proceed effectively in the hybrid strain.

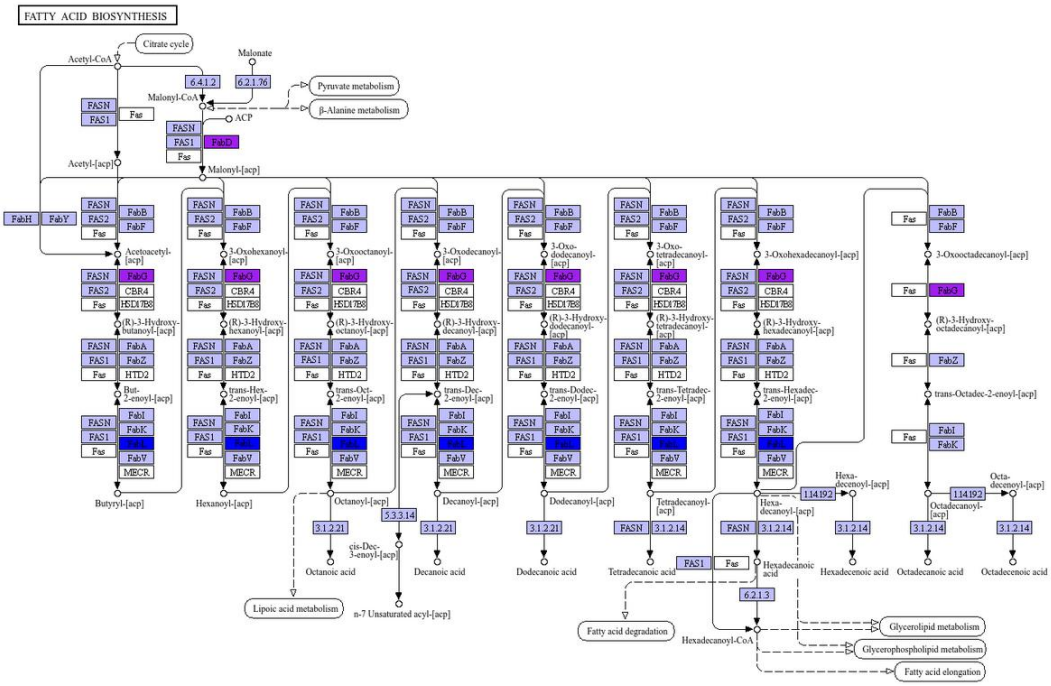


Figure 7. Fatty Acid Pathway with CBS 6016's enzymes part ½

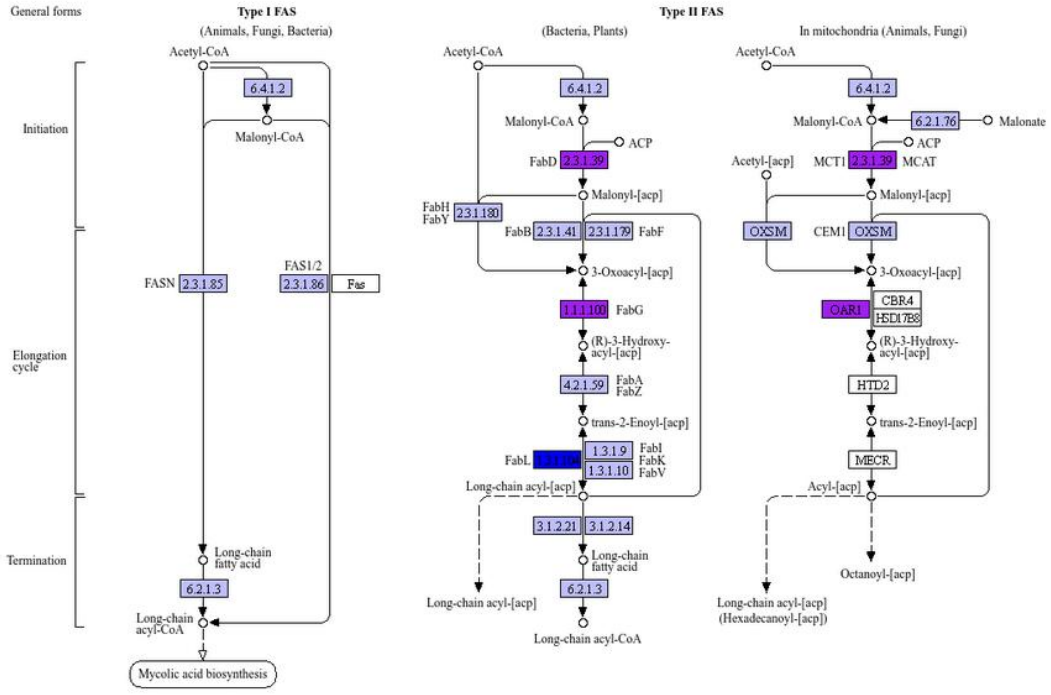


Figure 8. Fatty Acid Pathway with CBS 6016's enzymes part 2/2

Enzymes highlighted in green indicate potential losses during the hybridization event. These missing enzymes could influence the pathway's efficiency or

regulation, and further investigation would be needed to assess their impact on CBS 6016's lipid biosynthesis capabilities.

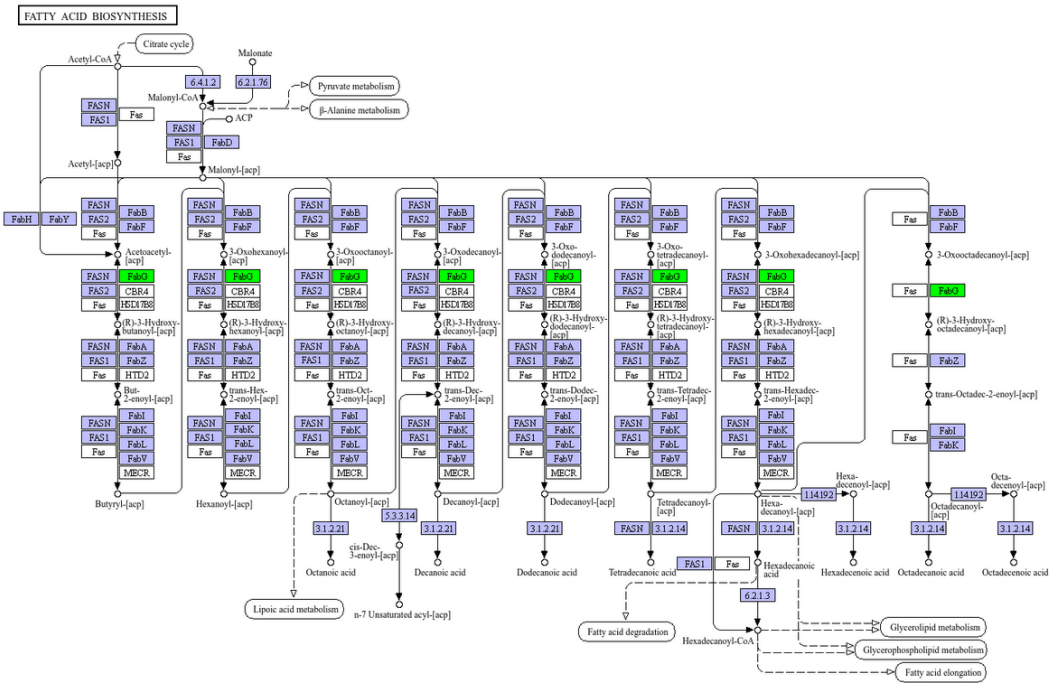


Figure 9. Fatty Acid Pathway with potential losses part ½

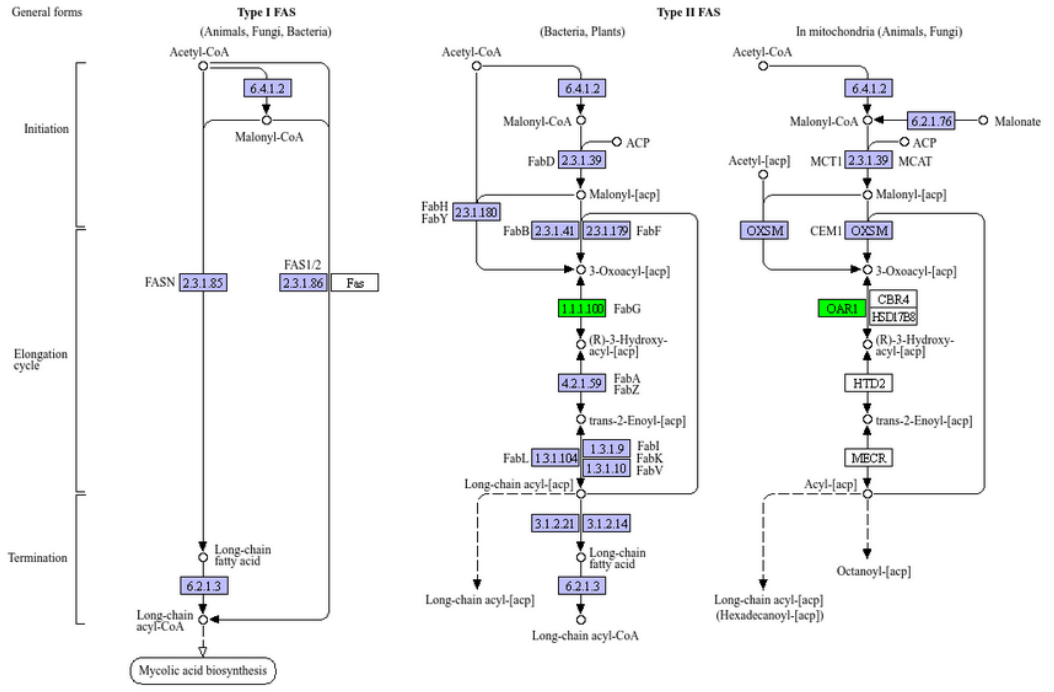


Figure 10. Fatty Acid Pathway with potential losses part 2/2

The pathway analysis indicates that CBS 6016 maintains a functional fatty acid biosynthesis pathway. While Figures 9 and 10 shows the absence of one FabG isoform (in green), a different isoform of FabG (See Figures 7 and 8) was transmitted by both parental strains. This redundancy likely reduces the impact of individual enzyme losses, ensuring the pathway remains functional despite these potential gaps.

Oxidative Phosphorylation

Cells depend on oxidative phosphorylation for the generation of energy, and so the efficiency of this pathway has a direct relation with the energy supply and metabolic activity of yeast.

The results obtained for this process shows a significant preservation of enzymes from both parental strains, as indicated by the predominant presence of purple highlights. This suggests that most of the enzymatic functions essential for this crucial energy-producing pathway were inherited from both CBS 349 and CBS 14, ensuring the maintenance of this vital process in the hybrid strain CBS 6016.

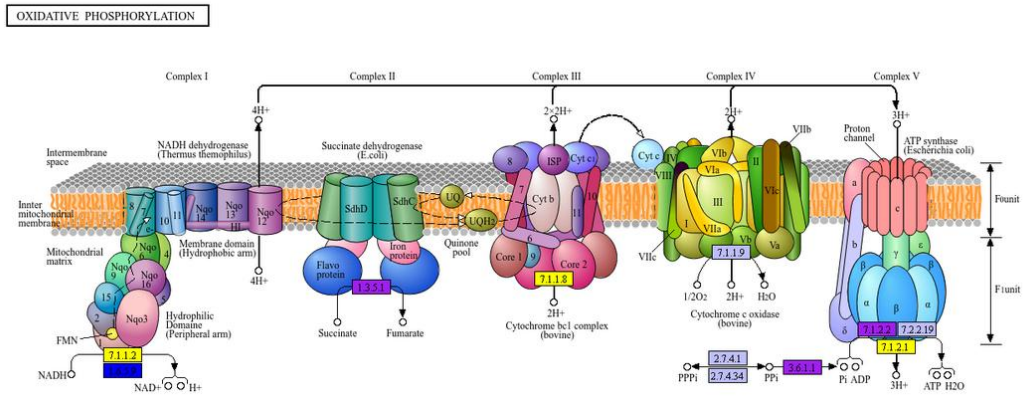


Figure 11. Oxidative Phosphorylation Pathway with CBS 6016's enzymes.

However, there is also evidence of potential enzymatic loss, as highlighted by the green mark at EC numbers 1.3.5.1 and 7.1.2.1, indicating that these specific enzymes may have been lost during the hybridization event. This loss could potentially impact the efficiency of the oxidative phosphorylation process in the hybrid strain, although the overall pathway remains largely intact due to the contributions from both parental strains.

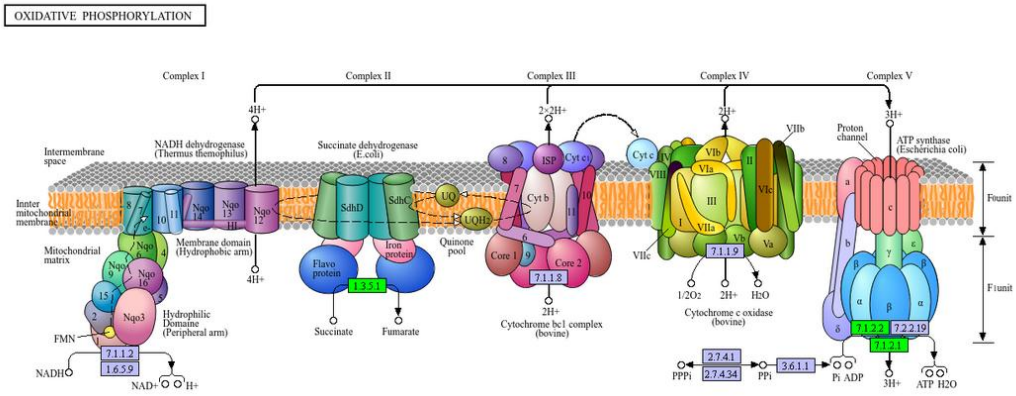


Figure 12. Oxidative Phosphorylation Pathway with potential losses after hybridization event.

Glycolysis / Gluconeogenesis

The glycolysis and gluconeogenesis pathways are essential for glucose metabolism, playing a crucial role in several metabolic activities. When we look into this pathway, we see that the majority of the EC numbers involved in this pathway have been preserved. Most of these enzymes appear to have been inherited from both parental strains, as indicated by the prevalence of purple highlights in the pathway. This suggests a strong retention of this critical metabolic pathway during the hybridization process. However, there is a notable exception: one EC number, highlighted in green, may have been lost during the hybridization event, indicating a potential gap in this metabolic function within the hybrid strain

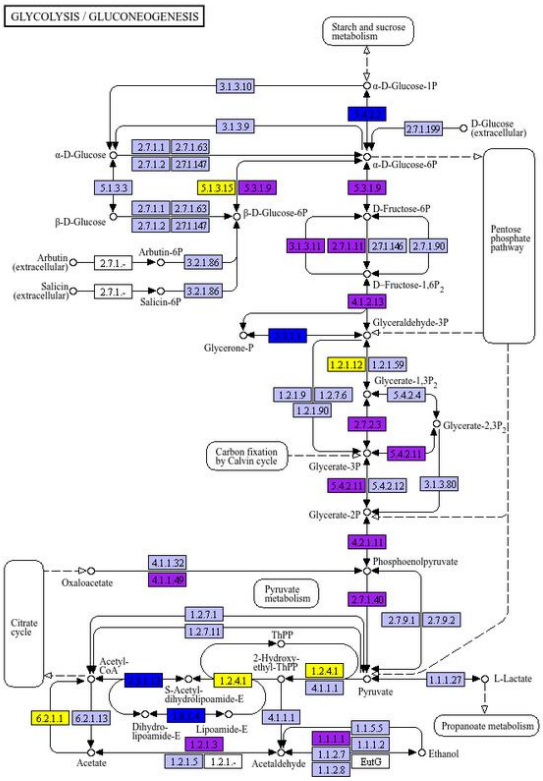


Figure 13. Glycolysis/Gluconeogenesis Pathway with CBS 6016's enzymes

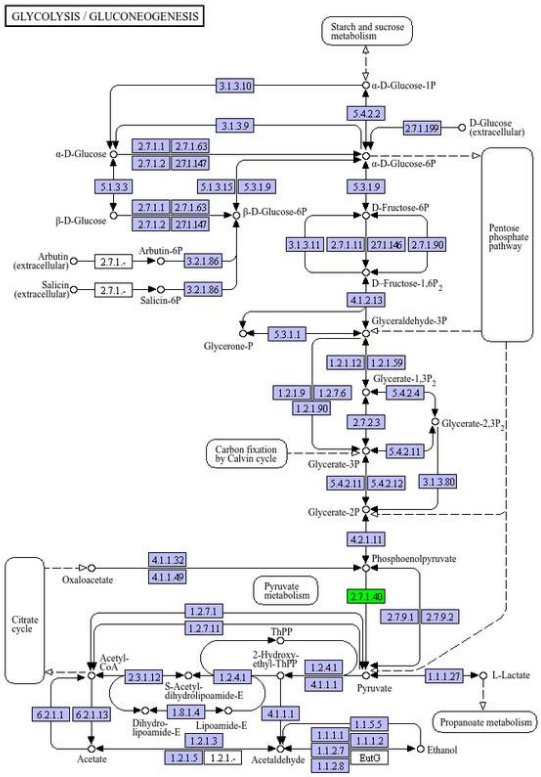


Figure 14. Glycolysis/Gluconeogenesis Pathway with potential losses after hybridization

From this, we can infer that CBS 6016 maintains effective metabolic processes essential for energy metabolism and cell survival. The presence of these crucial enzymes indicates a robust metabolic network in CBS 6016, capable of regulating glucose and energy efficiently.

TCA Cycle (Citrate Cycle)

The TCA cycle, also known as the citric acid cycle or Krebs cycle, is a fundamental metabolic pathway in living organisms, since it is responsible for generating energy and biosynthetic intermediates. This cycle has a series of reactions where carbohydrates, lipids, and proteins are oxidized to produce carbon dioxide, water, and energy in the form of ATP (adenosine triphosphate). The TCA cycle is vital for CBS 6016's energy metabolism, providing ATP, NADH, and FADH2 for cellular processes.

The analysis of this pathway shows a high level of conservation across the hybrid strain CBS 6016, with most of the EC numbers inherited from both parental strains, CBS 14 and CBS 349, as indicated by the predominance of purple highlights. This suggests that the core components of the TCA cycle, essential for energy production and intermediary metabolism, have been preserved through the hybridization process. The analysis also reveals potential losses, highlighted in

green, a single EC number appears to be lost, indicating that this specific enzyme might not have been transmitted during the hybridization event. This loss could have implications for the efficiency or regulation of the TCA cycle in CBS 6016, though further analysis would be required to fully understand the impact.

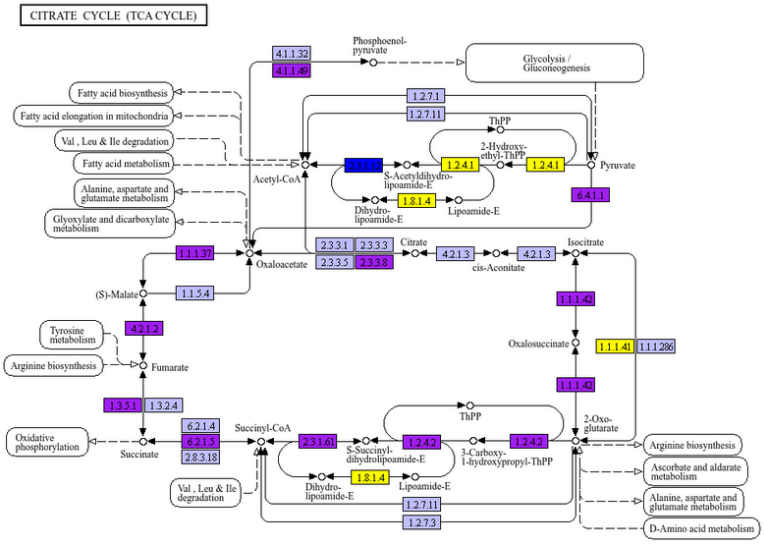


Figure 15. Citrate cycle (TCA cycle) with CBS 6016's enzymes.

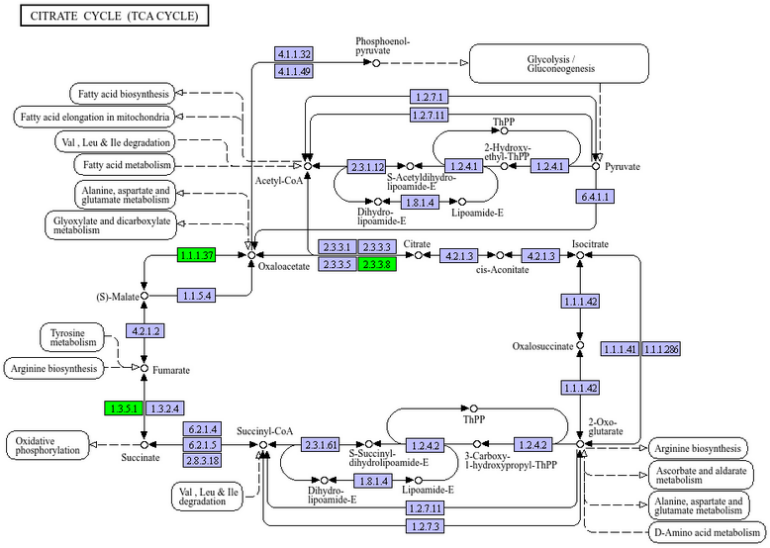


Figure 16. Citrate cycle (TCA cycle) with potential losses after hybridization

5. Discussion

The visualization of the complete genome of CBS 6016 (Figure 3), provides a view of the gene distribution patterns across all contigs when aligned with the parental strains CBS 14 and CBS 349.Across the entire genome, a predominant pattern emerges where certain contigs show a strong alignment with one parental strain over the other. This suggests that CBS 6016 is a mosaic of its parental strains, with distinct regions derived from either CBS 14 or CBS 349. The dense blue and red tracks indicate a successful alignment of CBS 6016's genome with its parents, reflecting the hybrid nature of this strain. However, gaps are visible where no alignment could be assigned, which could be indicative of areas where the CBS 6016 genome has diverged significantly from its parental origins, potentially due to recombination events, mutations, or even the presence of novel sequences that were not inherited from either parent. These gaps might also be a result of overly stringent filtering criteria, where legitimate alignments were excluded. This shows the importance of balancing the need for stringency with the need of capturing the full genetic complexity of hybrid organisms like CBS 6016.

To gain a deeper understanding of the gene distribution, we zoomed in on specific contigs such as contig 23, contig 29, and contig 19. By focusing on these contigs, we can observe more nuanced patterns that reflect the unique contributions of each parental strain. Contig 23 exhibits a predominance of CBS 14-origin genes, with a dense blue track across the contig. This suggests that this region of CBS 6016 is largely inherited from CBS 14, with relatively minor contributions from CBS 349. Contig 29 shows us a more mixed pattern, with contributions from both CBS 14 and CBS 349. The distribution of blue and red tracks suggests that this contig may have undergone recombination, resulting in a mosaic of genes from both parental strains. Finally, Contig 19 similarly to contig 23, this contig shows a strong alignment with CBS 14, with a number of regions dominated by blue tracks. The presence of some red segments indicates that CBS 349 also contributed to this region, but to a lesser degree.

Martin-Hernandez et al. (2023) observed that hybrid strains of *Rhodotorula toruloides* tend to inherit a substantial proportion of sequences from both parents, thereby maintaining essential metabolic functions. Our findings align with these observations, confirming that CBS 6016 has preserved crucial genetic material to support its metabolic activities (Martin-Hernandez et al., 2023).

KEGG Pathways Analysis

One of the challenges in accurately annotating and analyzing enzyme pathways is in the completeness in annotations for the Rhodotorula genus in databases like UniProt. Not all proteins are fully annotated with their corresponding EC numbers, which can lead to certain enzymes in the CBS 6016 genome not being assigned an EC number during the analysis. This can result in some EC numbers appearing uncolored on the KEGG pathway maps, which might suggest missing enzymes where, in reality, they may just lack proper annotation.

Another point is the fact that different proteins, including isoforms, can share the same EC number. This means that in some cases, a particular EC number might be marked as present (purple, red,blue or yellow) in one context, while also being indicated as potentially lost (green) in another context. This situation occurs when different isoforms of the same enzyme, inherited from each parent may be retained, while another may be lost during the hybridization process. At first glance, these color overlaps might suggest an error in the script or method used for the analysis. However, this actually highlights the nuanced nature of enzyme function and inheritance in hybrid organisms like CBS 6016. It underscores the importance of understanding the functional diversity within enzyme families and the role of different isoforms in maintaining metabolic pathways.

Fatty Acid Biosynthesis

The fatty acid biosynthesis pathway is very important for producing lipids, which are key components of cell membranes and are also used for energy storage. In *Rhodotorula toruloides*, this pathway is well-developed, which shows the yeast's strong ability to accumulate lipids, making it useful for industrial applications like making lipid-based products.

Our results show that CBS 6016, a hybrid strain of *Rhodotorula*, keeps a functional fatty acid biosynthesis pathway with enzymes inherited from both its parental strains, CBS 14 and CBS 349. The presence of several important enzymes from both parents ensures that the main steps in fatty acid elongation and termination are kept intact. This inheritance pattern is crucial for maintaining CBS 6016's ability to synthesize lipids, which agrees with earlier studies that highlight the importance of these pathways in yeasts that produce a lot of lipids (Zhu et al., 2012; Wen et al., 2020). Regarding the missing enzymes analysis, we observe that some might be missing, which is indicated by the green highlights in the pathway. These missing enzymes could suggest that during the hybridization process, some versions or parts of enzymes were not passed on, which might affect the overall efficiency of fatty acid biosynthesis. The lack of these enzymes could potentially disrupt the pathway, affecting lipid production or the regulation of the length of fatty acid chains, as mentioned in studies on enzyme functions in other *Rhodotorula* species (Zhu et al., 2012; Chen et al., 2015), however these same missing EC

numbers are also present on the analysis of the inherited enzymes, suggesting that a given EC number can have more than one specific peptide sequence.

To understand the impact of these potential losses on CBS 6016's metabolism and its ability to effectively synthesize lipids, further research is needed. Looking into how CBS 6016 might compensate for these missing enzymes, or understanding if there are other enzymes that can take over their function, could provide more insights into how flexible this strain's metabolism is.

These findings also highlight the complexity of enzyme inheritance in hybrid strains. The partial loss of some enzymes, alongside the retention of key ones, reflects the balance between keeping essential metabolic functions and the changes that occur during hybridization. As previous studies suggest, how these metabolic adjustments affect the strain could be important in determining how useful hybrid strains like CBS 6016 are for industrial applications (Wen et al., 2020; Adrio, 2017).

Oxidative Phosphorylation

Oxidative phosphorylation is an important mechanism for the production of ATP and regulation of metabolism in the cells and thus affects the energy supply. According to the present analysis of this pathway, enzymes of both CBS 349 and CBS 14 are conserved in CBS 6016, as demonstrated from the overall purple highlights dominating the pathway map. This means the majority of essential enzymatic activities that are need for the electron transport that is needed for ATP synthesis has been inherited from each of the parental strains and retained in the hybrid strain. However, the green highlights observed at EC numbers 1. 3. 5. 1 and 7. 1. 2. 1 are associated with possible enzymatic losses that may occur during the hybridization event. More concretely, these losses entail succinate dehydrogenase (EC 1. 3. 5. 1), and a subunit of NADH-ubiquinone oxidoreductase (EC 7. 1. 2. 1), enzymes that allow the transfer of electrons in the electron transport chain. The lack of these enzymes may result into weak oxidative phosphorylation and therefore may drawbacks the ATP generation and thus, the energy output in CBS 6016.

These could be due to the hybridization of the two strains, however, the vast part of the pathway is kept from the parental strains, and further contributions from each of the strain could partially makeup for the loss. This finding is in accordance with prior studies by Passoth et al. (2020) and Nagaraj et al. (2021) which highlighted the critical role of maintaining a robust electron transport chain for efficient ATP generation in hybrid strains. The findings here emphasize the importance of preserving key enzymatic functions during hybridization to sustain the metabolic efficiency of the organism.

Glycolysis / Gluconeogenesis

Another fundamental pathway for glucose metabolism that plays vital roles in energy production and other key metabolic processes. Our results indicate that most of the enzymes involved in these pathways have been preserved in the hybrid strain CBS 6016, as evidenced by the predominance of purple highlights in the KEGG pathway maps. This suggests that these critical metabolic functions were largely inherited from both parental strains, CBS 349 and CBS 14, ensuring the continuity of these essential processes in CBS 6016. This aligns with the findings of Martin-Hernandez et al. (2023), who emphasized the importance of these pathways in the metabolic fitness of Rhodotorula hybrids. Our analysis also revealed a potential loss, indicated by the green highlight on one EC number (2.7.1.40). This may suggest a gap in the hybrid strain's metabolic capabilities, although the overall pathway remains largely intact. The loss of this enzyme could impact the efficiency of glucose metabolism, but the preservation of other critical enzymes likely compensates for this deficiency, ensuring that CBS 6016 maintains robust metabolic functions.

TCA Cycle (Citrate Cycle)

This metabolic process plays a central role in the energy metabolism of cells by generating ATP, NADH, and FADH2, which are vital for various cellular processes. The analysis show that this cycle is mostly the same as in CBS 14 and CBS 349 and most enzymes are derived from both of these parental strains. Despite this loss, CBS 6016 has retained other oxidative necessary enzymes involved in TCA cycle including citrate synthase (EC 2. 3. 3. 1), isocitrate dehydrogenase (EC 1. 1. 1. This conservation is in concordance with the observation by Nagaraj et al, (2021) who emphasized the importance of maintaining TCA cycle enzymes for the metabolic flexibility of Rhodotorula strains.

On the other hand, the presence of a green-highlighted EC number in the case of the present analysis raised a possibility of loss of a particular enzyme during the process of hybridization. This could alter the regulation or even the speed of the TCA cycle in CBS 6016 and may affect the metabolic function of the latter.

Future Directions

Although CBS 6016 retains many crucial metabolic pathways, the partial loss of certain enzymes, particularly in fatty acid biosynthesis, oxidative phosphorylation, and the TCA cycle, it raises questions about its overall metabolic efficiency. Below are some possible areas for future research that could clarify these gaps and improve the use of CBS 6016:

- Investigate the Impact of missing enzymes on metabolism.

The potential loss of certain enzymes in important pathways like fatty acid biosynthesis, oxidative phosphorylation, and the TCA cycle suggests that CBS 6016's overall metabolism might be affected. Future research should aim to

understand how these missing enzymes influence the strain's metabolic efficiency, especially in terms of energy production and lipid synthesis.

- Explore how CBS 6016 adapts to missing enzymes

Although some enzymes are absent, CBS 6016 appears to maintain functional metabolic pathways. Research could focus on investigating whether other enzymes or alternative pathways compensate for these losses. Techniques like transcriptomics or proteomics could be used to identify these compensatory mechanisms.

- Improving the annotation of the CBS 6016 genome

Improving annotations, possibly through advanced bioinformatics tools or experimental methods. This would give a clearer picture of the strain’s metabolic capabilities and support future research efforts.

- Examine the role of isoforms.

Different isoforms of enzymes may help CBS 6016 adjust to the absence of certain enzymes, future research could delve deeper into these areas. Understanding how these isoforms function and contribute to the strain's overall metabolism could offer valuable insights into the metabolic flexibility of CBS 6016.

6. Conclusion

This thesis has explored the inheritance and functioning of key metabolic pathways in the hybrid strain CBS 6016. Through sequence alignment and pathway analysis, we found that CBS 6016 has kept many important metabolic capabilities from both of its parental strains, CBS 14 and CBS 349, which are essential for its survival and adaptability.

Our results show that CBS 6016 has inherited various protein sequences from both CBS 14 and CBS 349, which help to maintain crucial functions like fatty acid degradation, oxidative phosphorylation, glycolysis/gluconeogenesis, and the tricarboxylic acid (TCA) cycle. By keeping the important enzymes in these pathways, CBS 6016 is able to produce energy efficiently and carry out other vital metabolic processes, which are important for its viability and potential use in biotechnology. For example, the enzymes involved in fatty acid degradation suggest that CBS 6016 can perform lipid metabolism effectively. However, this ability might be slightly affected due to the partial loss of some butyryl-CoA dehydrogenase isoforms. Similarly, while the loss of a subunit of NADH-ubiquinone oxidoreductase could slightly impact energy production, the overall inheritance of the necessary components for ATP synthesis through oxidative phosphorylation shows that this strain has a strong potential for energy generation.

The strong retention of enzymes in glycolysis and the TCA cycle suggests that CBS 6016 is metabolically resilient and can perform the essential functions related to energy production, which is crucial for its growth, survival, and use in industrial applications, especially for biofuel and lipid production.

Comparing our findings with previous studies by Martin-Hernandez et al. (2023), Sànchez et al. (2022), Passoth et al. (2020), Nagaraj et al. (2021), and Buzzini et al. (2023) supports the idea that maintaining critical metabolic functions is important in hybrid strains.

The main goal of this thesis was to explore the genetic and metabolic capabilities of CBS 6016 by comparing its protein-coding sequences and metabolic pathway components with those of its parental strains, CBS 14 and CBS 349. The analysis revealed that CBS 6016 retains essential metabolic functions inherited from both parents, while also exhibiting partial loss of certain enzymes. This study offers valuable insights into the inheritance and functionality of key metabolic

pathways, providing a base for future research and biotechnological applications. It also points out areas where further exploration or optimization could enhance the industrial potential of CBS 6016.

References

Browning, S.R. and Browning, B.L. (2011). Haplotype phasing: existing methods and new developments. Nature Reviews Genetics, 12(10), pp.703–714. doi:https://doi.org/10.1038/nrg3054.

Buzzini, P., et al., 2023. The role of enzyme loss in metabolic reprogramming of hybrid yeasts. Biotechnology for Biofuels, 16(1), pp. 118-131. Available at: https://biotechnologyforbiofuels.biomedcentral.com/articles/10.1186/s13068-021-01916-y

Crandall, J.G., Fisher, K.J., Sato, T.K. and Chris Todd Hittinger (2023). Ploidy evolution in a wild yeast is linked to an interaction between cell type and metabolism. PLoS biology, 21(11), pp.e3001909–e3001909. doi:https://doi.org/10.1371/journal.pbio.3001909.

Gerstein, A.C. and Otto, S.P. (2009). Ploidy and the Causes of Genomic Evolution. Journal of Heredity, 100(5), pp.571–581. doi:https://doi.org/10.1093/jhered/esp057.

J. Carl Schultz, Cao, M. and Zhao, H. (2019). Development of a CRISPR/Cas9 system for high efficiency multiplexed gene deletion in Rhodosporidium toruloides. Biotechnology and bioengineering, 116(8), pp.2103–2109. doi:https://doi.org/10.1002/bit.27001.

Jiang, W., Li, C., Li, Y. and Peng, H. (2022). Metabolic Engineering Strategies for Improved Lipid Production and Cellular Physiological Responses in Yeast Saccharomyces cerevisiae. Journal of Fungi, [online] 8(5), p.427. doi:https://doi.org/10.3390/jof8050427.

Li, C.-J., Zhao, D., Li, B.-X., Zhang, N., Yan, J.-Y. and Zou, H.-T. (2020). Whole genome sequencing and comparative genomic analysis of oleaginous red yeast Sporobolomyces pararoseus NGR identifies candidate genes for biotechnological potential and ballistospores-shooting. BMC Genomics, 21(1). doi:https://doi.org/10.1186/s12864-020-6593-1.

Marçais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L. and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology, 14(1), p.e1005944. doi:https://doi.org/10.1371/journal.pcbi.1005944.

Martin-Hernandez, G.C., et al., 2023. Genome-scale metabolic models for understanding Rhodotorula toruloides metabolism. Yeast, 40(1), pp. 33-47. Available at: https://onlinelibrary.wiley.com/doi/10.1002/yea.3838

Moran, B.M., Payne, C., Langdon, Q., Powell, D.L., Brandvain, Y. and Schumer, M. (2021). The genomic consequences of hybridization. eLife, [online] 10, p.e69016. doi:https://doi.org/10.7554/eLife.69016.

Nagaraj, Y.N., et al., 2021. Biotechnological potential of Rhodotorula for biofuel production. Journal of Industrial Microbiology & Biotechnology, 48(11-12), pp. 1-12. Available at: https://www.sciencedirect.com/science/article/pii/S0888754321003694?via%3Dihub

Osman, M.E., Asharf Bakery Abdel-Razik, Zaki, K.I., Nesma Mamdouh and Heba El-Sayed (2022). Isolation, molecular identification of lipid-producing Rhodotorula diobovata: optimization of lipid accumulation for biodiesel production. Journal of Genetic Engineering and Biotechnology /Journal of Genetic Engineering and Biotechnology, 20(1), pp.32–32. doi:https://doi.org/10.1186/s43141-022-00304-9.

Pang, Y., Zhao, Y., Li, S., Zhao, Y., Li, J., Hu, Z., Zhang, C., Xiao, D. and Yu, A. (2019). Engineering the oleaginous yeast Yarrowia lipolytica to produce limonene from waste cooking oil. Biotechnology for Biofuels, 12(1). doi:https://doi.org/10.1186/s13068-019-1580-y.

Passoth, V., et al., 2020. The impact of hybridization on the metabolic capabilities of Rhodotorula strains. Fungal Biology and Biotechnology, 7(1), pp. 1-10. Available at: https://www.mdpi.com/2309-608X/8/4/323

Sànchez, C.C., et al., 2022. Lipid accumulation mechanisms in Rhodotorula toruloides. FEMS Yeast Research, 22(6), p. foac039. Available at: https://biotechnologyforbiofuels.biomedcentral.com/articles/10.1186/s13068-023-02294-3

Schultz, J.C., Mishra, S., Gaither, E., Mejia, A., Dinh, H., Maranas, C. and Zhao, H. (2022). Metabolic engineering of Rhodotorula toruloides IFO0880 improves C16 and C18 fatty alcohol production from synthetic media. Microbial Cell Factories, 21(1). doi:https://doi.org/10.1186/s12934-022-01750-3.

Weiß, C.L., Pais, M., Cano, L.M., Kamoun, S. and Burbano, H.A. (2018). nQuire: a statistical framework for ploidy estimation using next generation sequencing. BMC Bioinformatics, [online] 19, p.122. doi:https://doi.org/10.1186/s12859-018-2128-z.

Zhang, Y., Peng, J., Zhao, H. and Shi, S. (2021). Engineering oleaginous yeast Rhodotorula toruloides for overproduction of fatty acid ethyl esters. Biotechnology for Biofuels, 14(1). doi:https://doi.org/10.1186/s13068-021-01965-3.

Popular science summary

In our world today, the push for sustainable and environmentally friendly solutions is more important than ever. One area where this is especially true is in the production of lipids—fats and oils that are crucial for making biofuels, food additives, and even chemicals.

Some species or fungi called "oleaginous yeasts," can do much more. These yeasts can produce large amounts of lipids from low-value materials, such as agricultural waste or by-products that would otherwise be discarded. This ability makes them a potential game-changer for creating sustainable and eco-friendly alternatives to traditional lipid sources.

My project focuses on a specific type of yeast called *Rhodotorula*, which is particularly good at producing lipids. I studied three strains of *Rhodotorula* yeast—two parental strains and one hybrid strain.

My goal was to understand how this hybrid strain, CBS 6016, inherited the ability to produce lipids and how well it can perform this task. To do this, I looked at the genetic makeup of CBS 6016 and compared it to its parent strains. By examining key metabolic pathways—the processes by which the yeast converts raw materials into energy and lipids—I was able to see which enzymes were passed down from each parent.

CBS 6016 inherited many of the important enzymes needed for lipid production from both parent strains. This means that it should be able to efficiently produce lipids, making it a strong candidate for use in sustainable biotechnological applications. However, I also found that some enzymes were missing in the hybrid strain, which could slightly reduce its efficiency.

In conclusion, my research shows that CBS 6016 is a robust hybrid yeast strain with strong potential for producing lipids sustainably. It offers a glimpse into how we might use yeast to create greener alternatives to traditional lipid production, reducing our reliance on fossil fuels and minimizing environmental impact.

Acknowledgements

First, I would like to thank my family. Your support, love, and encouragement have been the backbone of my journey through this academic work. You believed in me even when I wasn't sure of myself, and I am very grateful for that.

To my friends, thank you for being there for me. Your understanding and companionship have made this journey easier, and your presence has always reminded me that I'm not alone in this.

I am also very thankful to my teachers, who have guided me and shared their knowledge throughout my studies. Your dedication to teaching has inspired me to keep learning and growing.

A special thanks goes to my supervisor, Bettina Mueller. Your guidance and support throughout this research have been invaluable. You pushed me to think critically and helped me develop as a researcher. I couldn't have done this without your help.

Finally, I want to thank everyone who has helped me along the way. Whether through academic advice, moral support, or just a kind word, your contributions have made a difference, and I am truly grateful.

Thank you all.

Publishing and archiving

Approved students' theses at SLU are published electronically. As a student, you have the copyright to your own work and need to approve the electronic publishing. If you check the box for **YES**, the full text (pdf file) and metadata will be visible and searchable online. If you check the box for **NO**, only the metadata and the abstract will be visible and searchable online. Nevertheless, when the document is uploaded it will still be archived as a digital file.

If you are more than one author you all need to agree on a decision. Read about SLU's publishing agreement here: https://www.slu.se/en/subweb/library/publish-and-analyse/register-and-publish/agreement-for-publishing/.

☒ YES, I/we hereby give permission to publish the present thesis in accordance with the SLU agreement regarding the transfer of the right to publish a work.

☐ NO, I/we do not give permission to publish the present work. The work will still be archived and its metadata and abstract will be visible and searchable.