UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
DEPARTAMENTO DE BIOLOGIA CELULAR E MOLECULAR

Synthetic biology applied to *Rhodosporidium toruloides* for fine chemical production

LUÍSA CZAMANSKI NORA

Ribeirão Preto
Brasil
2022

LUÍSA CZAMANSKI NORA

Synthetic biology applied to *Rhodosporidium toruloides* for fine chemical production

Tese apresentada à faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para a obtenção do título de Doutora em Ciências – Área de concentração: Biologia Celular e Molecular

Orientador: Dr. Ricardo Roberto da Silva

Versão corrigida. A versão original encontra-se disponível tanto na Biblioteca da Unidade que aloja o Programa, quanto na Biblioteca Digital de Teses e Dissertações da USP (BDTD)

Ribeirão Preto
Brasil
2022

NORA, Luísa Czamanski

Biologia sintética aplicada a *Rhodosporidium toruloides* para produção de químicos finos, 2022.

150 f. : il. ; 30 cm

Tese de Doutorado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Biologia Celular e Molecular.

Orientador: Dr. Ricardo Roberto da Silva.

1. Biologia sintética 2. Engenharia metabólica 3. Transcriptômica 4. Fatores de transcrição 5. Cana-de-açúcar 6. Fungos 7. Monoterpeno 8. Pineno 9. Linalol 10. Geraniol

Aluna: **Luísa Czamanski Nora**

Título: **Synthetic biology applied to *Rhodosporidium toruloides* for fine chemical production**

Tese apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo para obtenção do título de Doutora em Ciências. Área de concentração: Biologia Celular e Molecular. Orientador: Dr. Ricardo Roberto da Silva

Aprovado em: ___/___/___

Banca Examinadora

Prof(a).Dr(a).:_____
Instituição:_____
Assinatura:_____

Prof(a).Dr(a).:_____
Instituição:_____
Assinatura:_____

Prof(a).Dr(a).:_____
Instituição:_____
Assinatura:_____

Dedico essa tese à minha família, por serem minha força e inspiração para tudo, e por sempre torcerem pelas minhas conquistas, mesmo que isso signifique estar fisicamente longe.

AGRADECIMENTOS

Quando me diziam que um projeto de doutorado nunca segue uma trajetória linear, eu não fazia ideia de quão sinuosa e inusitada ela poderia ser. Agora eu entendo que são essas curvas que nos fazem crescer como pessoas e profissionais, e eu só tenho a agradecer a todos que de alguma forma me ajudaram a cruzar esse caminho tortuoso. Aqui, queria deixar registrado alguns desses agradecimentos. Ao meu orientador, Dr. Ricardo Roberto da Silva, por ter aceitado o desafio de me orientar no último ano do doutorado e ter contribuído tanto mesmo de longe. Ao meu antigo orientador, Dr. Rafael Silva Rocha, pelas oportunidades e ensinamentos. Ao suporte financeiro da Fundação de Amparo a Pesquisa do Estado de São Paulo através do processo FAPESP 2019/04942-7, do CNPq através do Processo CNPq 140212/2019-1 e da CAPES. Ao Departamento de Biologia Celular e Molecular da FMRP-USP e a todos os seus funcionários pelo apoio. Ao Supera Parque da USP pela oportunidade de trabalhar no diagnóstico de COVID-19 em 2020 e a todos os voluntários do projeto SuperAção por terem sido uma faísca de esperança em um momento tão difícil. A todos os colegas que passaram pelo saudoso laboratório de Biologia Sintética da FMRP-USP pela amizade que ultrapassa tempo, distâncias e fronteiras. Em especial à Ananda, por ter divido comigo a casa, os sonhos, os desafios, e ser meu ombro amigo em qualquer lugar do mundo. Ao Tiago, pela paciência, conselhos, conversas e incentivo. E ao Murilo, por ter me ajudado a *encontrar motivos* para continuar (com o perdão do trocadilho).

ACKNOWLEDGMENTS

Words fail to express how grateful I am to Dr. Nemailla Bonturi and Dr. Petri-Jaan Lahtvee for inviting me to come to Estonia to work on our favorite yeast, which not only changed the direction of my PhD project, but also of my entire life. Thank you for your trust, your guidance and for all the opportunities you have been giving me. I'm also grateful to Tallinn University of Technology and Dora Plus Scholarship for the financial and structural support, to my BIOCEB students, George, Gustavo, Belén and Aryian, for their outstanding work with the geraniol and linalool constructs and to my colleagues from the Bioengineering Lab at TalTech for making everyday life so joyful. Thanks to all the collaborators from around the world, including the Joint BioEnergy Institute, Imperial College of London and DTU Biosutain for making this work possible, and to Dr. James Kirby for always being my guide. A very special thanks to all the amazing friends I made in Estonia, who have helped me in this crazy journey in so many different ways. And to Ville, for brightening up even the darkest Estonian days.

“Trust those who seek the truth, doubt those who find it.”

– André Gide

GENERAL INDEX

RESUMO

NORA, Luísa Czamanski. Biologia sintética aplicada a *Rhodosporidium toruloides* para produção de químicos finos. [tese]. Ribeirão Preto: Universidade de São Paulo, Faculdade de Medicina de Ribeirão Preto; 2022.

O aumento da população mundial e dos padrões de vida está afetando os recursos do planeta. O conceito de bioeconomia circular estimula o uso de recursos biológicos renováveis para gerar produtos de base biológica, de forma a reduzir o uso de combustíveis fósseis e criar um modo de vida mais sustentável. Neste contexto, microrganismos que podem produzir químicos finos a partir de matéria-prima renovável estão em alta demanda para aplicações em biorrefinarias. A combinação de estratégias de biologia sintética, bioinformática e engenharia metabólica permite a otimização de organismos hospedeiros com potencial para se tornarem fábricas microbianas. A levedura não convencional *Rhodosporidium toruloides* é um dos microrganismos com grande potencial para ser aplicada para este propósito, pois é capaz de crescer em uma ampla gama de substratos e suportar alguns dos estresses causados por bioprocessos. Portanto, este trabalho buscou entender os comportamentos transcricionais de *R. toruloides* cultivado em condições de estresse e utilizando caldo de cana-de-açúcar como substrato, através do sequenciamento de RNA. Um *pipeline* de bioinformática foi desenvolvido, visando a descoberta de novos elementos *cis*-regulatórios a partir dos dados de transcriptômica, e essa ferramenta pode ser aplicada a outros hospedeiros microbianos no futuro. Este trabalho também se aventurou na produção de químicos finos, tendo *R. toruloides* como hospedeiro e aplicando técnicas de engenharia

metabólica e métodos de *assembly* de última geração. Três importantes terpenos foram escolhidos: pineno, linalol e geraniol. Embora os terpenos não tenham sido detectados, o ciclo de "*Design, Build, Test, Learn*" foi aplicado com sucesso para revelar o que pode ser aprimorado em tentativas futuras. Este trabalho demonstra métodos para a interpretação de dados de transcriptômica, a detecção de elementos regulatórios em organismos não convencionais e a engenharia de microrganismos para produzir químicos finos. Essas estratégias ainda podem ser otimizadas para criar fábricas microbianas para produção renovável de químicos finos, trazendo nosso mundo para uma realidade mais sustentável.

Palavras-chave: Biologia sintética, engenharia metabólica, transcriptômica, fungos, fatores de transcrição, cana-de-açúcar, monoterpeno, pineno, geraniol, linalol.

ABSTRACT

NORA, Luísa Czamanski. Synthetic biology applied to *Rhodosporidium toruloides* for fine chemical production. [tese]. Ribeirão Preto: Universidade de São Paulo, Faculdade de Medicina de Ribeirão Preto; 2022.

The current increase in world population and standard of living is taking its toll on the planet's resources. The circular bioeconomy concept comes to spark the use of renewable biological resources to generate bio-based products, in order to reduce the use of fossil fuels and create a more sustainable way of living. In this context, microorganisms that can produce highly valuable chemicals using renewable feedstock are in high demand for applications in biorefineries. The combination of synthetic biology, bioinformatics and metabolic engineering strategies allows the optimization of hosts that can become microbial cell factories. The non-conventional yeast *Rhodosporidium toruloides* is one of the microorganisms with great potential to be applied for this purpose, since it is able to grow in a wide range of substrates and to withstand some of the stresses caused by bioprocesses. Thus, this work sought to understand transcriptional behaviors of *R. toruloides* when growing in stress-related conditions and with sugarcane as substrate, using RNA sequencing. A bioinformatic pipeline was then developed aiming the discovery of novel *cis*-regulatory elements from the yeast transcriptomic data, a tool that can be applied to other microbial hosts in the future. This work also ventured at producing valuable chemicals using *R. toruloides* as a host, applying metabolic engineering techniques and using state-of-the-art assembly methods. Three important terpenes were chosen for this production: pinene, linalool and geraniol. Although they

were not detected, the Design, Build, Test, Learn cycle was successfully applied to reveal what can be improved in future endeavors. This work demonstrates methods for the interpretation of transcriptomic data, the detection of regulatory elements in non-conventional organisms, and the engineering of hosts to produce valuable chemicals. These strategies can be further optimized to create microbial cell factories for production of valuable chemicals in a green, renewable way, bringing our world to a more sustainable reality.

Keywords: Synthetic biology, metabolic engineering, transcriptomics, fungi, transcription factors, sugarcane, monoterpene, pinene, geraniol, linalool.

I. INTRODUCTION

1. The circular bioeconomy

Our current market-oriented economic system is not working. Over 70% of the world's population is being affected by rising inequalities, a third of the land is severely degraded, and we are losing forests and animals at an alarming rate (World Economic Forum, 2020). Energy demand is increasing worldwide; therefore, fossil-fuel-based resources are finishing and greenhouse gas emissions are worsening, aggravating environmental problems like global warming (Devi *et al.*, 2022).

In contrast to the current linear model defined by "build, consume and waste," a circular economy model emphasizes economic growth based on activities that are distinguished from the consumption of finite resources to achieve positive benefits for the whole society (Y. Liu *et al.*, 2021). According to the Ellen MacArthur Foundation (The Ellen Macarthur Foundation, 2022), a circular economic model aims to create financial, natural, and social capital without waste, keeping products useful while regenerating biological systems. It highlights the need to develop business models in ways that disconnect business prosperity from the mere consumption of products. It's also about making products that can be easily reused and recycled, minimizing waste and maximizing their value throughout their lifecycle. The circularity concept is precisely based on this "4R" framework: reduce, reuse, recycle, and recover (Tan and Lamers, 2021).

Similar to the circular economy, the bioeconomy is an emerging field whose definitions include all services and sectors that produce or process biomass or use biological resources in any way, utilizing renewable resources to produce food, materials and energy, in the pursuit of an industrial transition of final products for economic,

environmental, social, and national security (Liobikiene *et al.*, 2020). Hence, the bioeconomy involves using renewable biological resources such as biomass to produce bio-based products that are considered renewable and potentially offer reduced environmental impacts compared to their petroleum-derived counterparts. Pursuing this framework not only means replacing fossil energy with renewable energy, but it also means moving to fossil-free materials, substituting carbon-intense products like plastics, concrete, steel, and synthetic textiles for lower carbon alternatives (Palahí and Adams, 2020).

Independent of the concept, carbon will remain the core molecule that builds fuels, fiber, chemicals, and other products. The main advantage of bio-based products is that they substitute fossil carbon with biogenic carbon originating from biomass, so their carbon dioxide (CO_2) emissions are also biogenic, which is considered carbon-neutral. Additionally, when used as an energy source, biofuels and biopower are renewable energy that can displace fossil fuels and power to alleviate fossil resource depletion (Tan and Lamers, 2021; Dutta *et al.*, 2022).

1.1. Turning biomass into fine chemicals as part of circular bioeconomy

The search for alternative sources for energy and chemical generation in a way that matches the circular bioeconomy framework is turning biorefineries into essential needs (Devi *et al.*, 2022). This is due to the fact that biorefineries can use different kinds of biomass as a source of raw materials, green energy and fossil fuel replacements (Dutta *et al.*, 2022). Lignocellulosic biomass is an excellent carbon source for biorefineries that

produce high-volume biofuels and bioproducts; for being the most abundant renewable carbon source after CO_2 and is widely available (Geiselman *et al.*, 2020).

Several different methodologies can be applied for the valorization of lignocellulosic biomass and conversion into high-added-value products. Some examples include mechanical processes, such as pelletization – a multi-step process for producing energy carriers – and thermo-chemical, such as combustion, gasification and pyrolysis. But the processes mainly utilized in biorefineries are biochemical ones based on biochemical processes, such as microbial conversion (Yaashikaa, Senthil Kumar and Varjani, 2022). Microbial conversion can turn several different kinds of biomass into valuable products, including agro-industrial residues. The differences in biomass source and the configuration of the biorefinery will be based on local factors, such as agricultural practices, climate, feedstock availability, and transportation (Dutta *et al.*, 2022). When microorganisms are applied to the conversion to chemicals and other products, competition with food production or land resources are reduced and fluctuations due to changes in climate or environment are avoided (Bonturi, Pinheiro, Monteiro de Oliveira, *et al.*, 2022), so one can choose whichever feedstock is more readily available for that specific region. Feedstocks can be made accessible for microbial conversion through hydrolysis, and typical examples are hydrolysates from rice, wheat and barley straw, coconut husk, corn stalk, sorghum, wood, sugarcane plants and sugarcane bagasse (Devi *et al.*, 2022). For instance, Brazil is prominent in sugarcane production, reaching 4.6 million hectares cultivated and more than 600 million tons obtained at harvest (Soccol *et al.*, 2017).

Lignocellulosic biomass comprises two types of polymers: carbohydrate polymers (cellulose and hemicellulose) and aromatic polymers (lignin). The composition of these parts is as follows:

a) Cellulose - a polymer of hexoses (C6 sugars) that comprises from 30 to 60% of the lignocellulosic mass; the repeating unit in cellulose is disaccharide cellobiose, consisting of two glucose molecules.
b) Hemicellulose - second main polymer in lignocellulose, about 20 to 40% of the total mass, composed of several heteropolymers, such as xylan, galactomannan, arabinoxylan and xyloglucan, of which the repeating units are pentoses (C5 sugars), mainly xylose. It has an amorphous structure that appears branched with functional groups like acetyl, methyl, glucuronic acid and galacturonic acid.
c) Lignin - a complex structure, comprising 15 to 20% of lignocellulose, is a non-carbohydrate polymer presenting phenylpropanoid units and phenolic compounds.

Because of its complex, non-uniform structure, degrading lignocellulose is not an easy task (Bonturi *et al.*, 2017; Devi *et al.*, 2022). To be employed in bioprocesses, it usually needs some form of pre-treatment or detoxification. Different methods of pre-treatment were developed, and other feedstock might require different types of treatment. Either way, this detoxification plays a pivotal role in the efficient breakdown of lignocellulosic biomass. (Wagle *et al.*, 2022) recently reviewed the whole variety of pre-

treatments that can be applied to different sources of biomass. These processes help to solubilize and hydrolyze the sugars, making them readily available for the microorganisms. However, they also release undesired product molecules known as inhibitors of microbial growth: acetic acid, furfural, 5-hydroxymethylfurfural (HMF), and phenolic compounds (Bonturi *et al.*, 2017; Wagle *et al.*, 2022). The microorganisms suitable for microbial conversion need to consume most of those sugars and survive the harsh environment caused by the accumulation of toxic compounds. Thus, for this microbial conversion to occur, robust microorganisms need to be employed to hydrolyze the lignocellulosic biomass.

Currently, more time and cost-efficient strategies are required for efficient deconstruction, detoxification and conversion of the different types of biomasses into consumable products. To do that, new technologies need to be developed, not only to refine the pre-treatment of the chosen feedstock, but also to improve the microorganisms used to digest, hydrolyze, and convert those substrates. As discussed, the organisms need to be able to successfully convert the sugars into the desired compound in competitive amounts, while withstanding the harsh environments triggered by biorefineries: high-temperature conditions, contaminations, and accumulation of toxic compounds in the media, among others. This optimization needs to rely on the fields of synthetic biology and metabolic engineering for creating robust strains and reaching high yields of biofuel and biochemical production using renewable substrates.

2. Synthetic biology for tool development

The development of recombinant DNA technologies, DNA cloning and transformation of cells unlocked a whole new world of possibilities for the molecular biology field, providing new perspectives on how researchers explore and engineer living systems (Nora *et al.*, 2018). From the improvement of these manipulation techniques, new automated DNA synthesis, and error checking and assembly of large DNA fragments, there was the creation of the field of synthetic biology. This field of study emerged about 20 years ago in the pursuit of turning living organisms into controllable systems, in analogy to mechanical or electronic systems. Synthetic biology stands out from other engineering disciplines in its approach and research goal. The main idea is to treat living organisms like computers, so every system working inside this living organism should be able to be programmed, just like computational systems. To do that, standardization and characterization of biological parts are one of the main focuses of this emerging science. It is a hybrid discipline, combining elements of engineering in basic science, formulating new rules for engineering organisms and achieving the desired goals (Nora, Westmann, *et al.*, 2019).

A critical concept in the synthetic biology framework is the use of the Design, Build, Test, Learn (DBTL) cycles **(Figure 1)**. The whole notion comes from understanding the biological system that needs to be engineered, creating new behaviors and then studying them - to understand how this can be improved even further. The "design" part comes from seeking to understand the desired host, or "chassis" - what is commonly called in this area of study. Also, to understand which genetic parts and biochemical pathways are the targets. With all this information gathered, it is possible to properly design

combinations of genes into genetic circuits and proceed to the next steps of the cycle. The "build" stage is for creating new expression *cassettes*, or genetic circuits, using state-of-the-art cloning technology, such as Golden Gate or Gibson assembly. The "test" part of the cycle requires the host organism, the "chassis," to express the circuits in specific environments for the emergent behaviors to be characterized. Lastly, the "learn" stage consists of understanding what is functional and what is not, to improve the circuits for the subsequent cycle iterations. This cycle can last as many iterations as needed until the desired properties are achieved (Nielsen and Keasling, 2016; Opgenorth *et al.*, 2019).

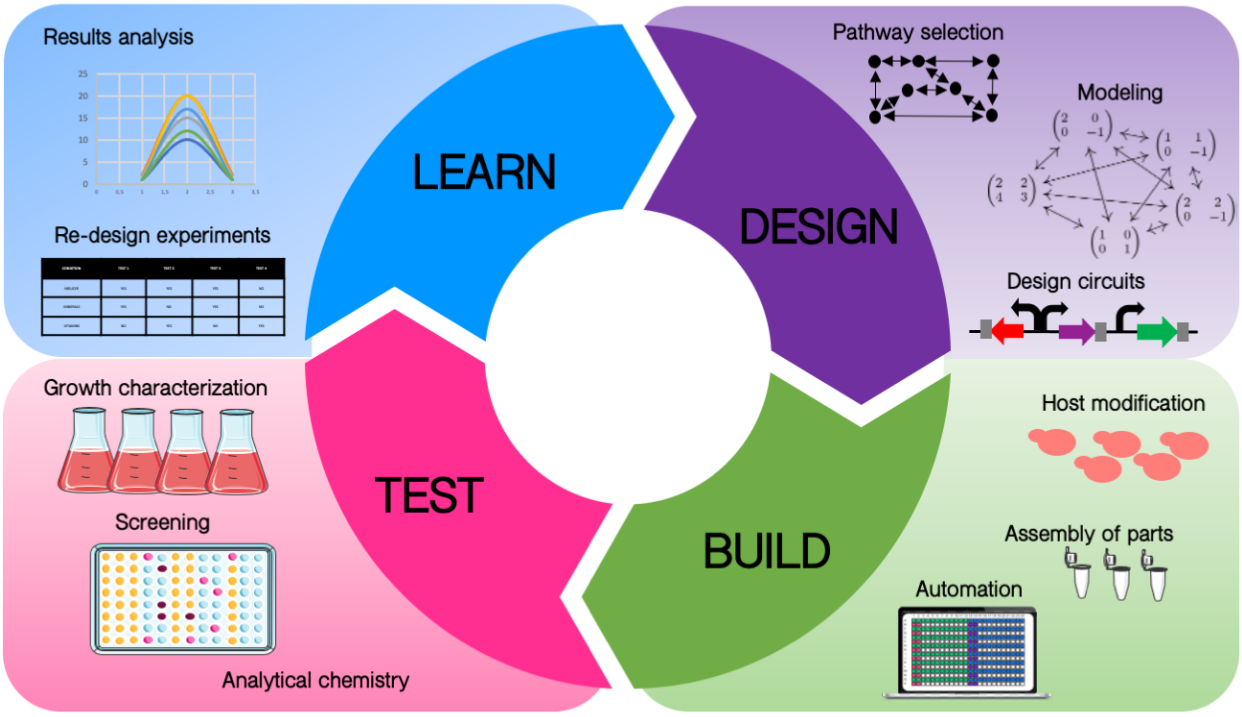


Figure 1. Scheme representing the steps in each **Design – Build – Test – Learn** cycle applied to the synthetic biology framework.

Modularity and orthogonality are also essential concepts in synthetic biology, where genetic parts are easily moved from one circuit to another for fine-tuning the genetic circuits using the same state-of-the-art methodologies mentioned earlier. They can then be combined in different contexts and still provide reliable and predictable behaviors, working similarly in different hosts, simulating what is expected for computational systems (de Lorenzo, 2011).

2.1. Genetic parts to develop toolboxes

In order to achieve the aforementioned modularity and orthogonality, synthetic biologists look for the characterization of the genetic parts in insulated environments. Due to this, scientists have a better understanding of how all the components of their systems work even before the project starts - and every part of the system matters. To this end, the genetic elements are usually cataloged by their specific functions. Several components can constitute and play important roles in genetic circuits, including the genes of interest, promoters, terminators, selection markers, transcription regulators, inducers, repressors, and enhancers, not to mention the different types of plasmid backbones and the parts that compose it (Nora *et al.*, 2018). Here, we will describe some of the main players in genetic circuit regulation: promoters, transcription factors, and terminators.

2.1.1. Promoters

A major need on the tool's front is efficient and stable promoters expressing genes of interest and selection markers. Different promoters can modulate not only when and if our gene of interest will be expressed, but also the level of expression. We can build our expression system with a strong promoter, resulting in high expression, or a weak promoter, resulting in low expression of the gene of interest. The second option is more relevant when the resulting product of expression is toxic to the cell, for example, or even to avoid metabolic burden. Furthermore, another way to control gene expression is by using inducible promoters. These are promoters that need specific molecules or signals to be activated, and these activators can be added to the culture medium as needed. While promoters that are responsive to certain metabolites are valuable tools, a toolset of well-characterized constitutive promoters remains necessary to explore the full potential of strain engineering. In this context, functional bidirectional promoters are also a valuable tool, since they can be a solution for efficient gene co-expression (Vogl, Kickenweiz, *et al.*, 2018; Nora, Wehrs, Kim, Skerker, *et al.*, 2019).

2.1.2. Transcription factors

Transcription factors (TF) and transcription factors binding sites (TFBS) are essential for tool development in synthetic biology because they can directly or indirectly drive different gene expression patterns. Thus, identifying those proteins and their binding sequences is critical for better transcription control. Finding new TF motifs also improves the understanding of gene regulation in the cell as a whole, which can serve as a basis for new bioengineering applications (Martins-Santana *et al.*, 2018). However, finding

those motifs is a difficult task, since the primary nucleotide sequence is one of many characteristics that should be taken into consideration. Some other factors, such as DNA shape and genomic context, can influence nucleotide preferences by the TFs (Inukai, Kock and Bulyk, 2017). With the help of computational approaches, TFBSs can, for example, be represented as position weight matrices (PWMs) in an attempt to define the binding energy of the DNA-protein interaction statistically depending on the data type that originated it.

2.1.3. Terminators

Terminator sequences are essential to molecular regulatory modules in a synthetic biology framework. They are important in the final step of transcription for the generation of stable mRNA molecules and as insulators in genetic circuits (Curran *et al.*, 2013; Song *et al.*, 2016). Still, although much is known about the regulatory elements that start the transcription, considerably less information is available about transcriptional termination (Amarelle *et al.*, 2018). Some early studies have shown how synthetic and minimal terminators that can influence transcription and mRNA half-life (Guo and Sherman, 1996), and, more recently, further studies have demonstrated the use of engineered terminators for synthetic biology applications (Ito *et al.*, 2016; Martins-Santana *et al.*, 2018).

2.2. State-of-the-art assembly methods

For all these parts to be combined in a functional circuit, they must be assembled in constructs that can be delivered to the host cells. The first cloning method available was established in 1973 when the first plasmid backbone had its resistance marker exchanged (Cohen *et al.*, 1973). This method relied on digestion by restriction enzymes and ligation by DNA ligases. It was a very successful methodology, since it allowed most of the advances in all the molecular biology-related fields up until this day. However, it is very laborious and time-consuming work and, to reach the synthetic biology goal of automatizing circuit building, novel techniques must be developed. Several assembly methods have been developed in the past 40 years, all with various advantages and disadvantages (Nora *et al.*, 2018). Some examples are such as Sequence and Ligation-Independent Cloning (SLIC) (Li and Elledge, 2007), Gateway Recombination Cloning (Reece-Hoyes and Walhout, 2018) and Circular Polymerase Extension Cloning (CPEC) (Quan and Tian, 2009). Yet, the two most important ones regarding the modularity, versatility, and suitability for automated processes, including robotic ones, are Gibson assembly and Golden Gate Assembly.

2.2.1. Gibson Assembly

Gibson Assembly was developed in 2009 when scientists were trying to create the first fully synthetic bacterial genome (Gibson *et al.*, 2009). It consists of a one-pot reaction containing three enzymes: one exonuclease, one polymerase and one ligase. That specific exonuclease cleaves DNA from the 5' end to the 3' end, leaving DNA overhangs in all sequences. Then, homologous regions hybridize, and the ligase binds

them together. The polymerase fills all the gaps remaining in this reaction, and fully built plasmids come out of the assembly in only one hour. To have the constructs ready for this type of assembly, primers must be synthesized containing homology arms for each part of the construct. The main advantage of this method is not requiring specific sequences at the end of each component so that any part can be assembled without restriction, without the need for specific restriction enzyme recognition sites.

2.2.2. Golden Gate Assembly

Golden Gate Assembly is the most beloved methodology of DNA assembly in the synthetic biology world, owing to the fact that it is the most modular of all the methods created up until this day, adaptable to automated processes. The technology is based on the activity of Type IIS restriction enzymes and T4 DNA ligase, with either sequential or simultaneous activity of both enzymes. Type IIS restriction endonucleases cleave the DNA outside the recognition sequence, therefore, not leaving any scars in the resulting construct. The recognition site for the endonucleases is created with a 4-nucleotide sequence overhang, which can be standardized for specific toolkits and libraries. This standardization of recognition site position allows easy exchange of parts. For instance, if all promoters in a promoter library have the same overhangs, they can be easily shuffled around in one single reaction. This method makes the diversification of parts when building constructs a less laborious process (Weber *et al.*, 2011; Bonturi, Pinheiro, Monteiro de Oliveira, *et al.*, 2022).

2.3. Bioinformatics

Bioinformatics is a field of study that began even before the human genome consortium started in the 1990s, but it has constantly been evolving ever since (Gauthier *et al.*, 2019). With computational methods being applied to Biology, analyzing information from metabolism, genomes and transcriptomes from a wide range of organisms was made possible. Bioinformatics integrates with the synthetic biology framework during all DBTL cycles (Figure 1). All the steps in the "design" of experiments are aided by computational tools, e. g. generating genome-scale metabolic models to simulate biochemical pathways in order to decide which genes to alter. Moreover, they allow the integration of information from different genomes, metabolomes and transcriptomes, to discover new biological parts that can be further tested in the desired chassis (Martins-Santana *et al.*, 2018). Several tools were created throughout the years to help with genetic circuit design, and even DNA assembly planning is entirely done computationally first. Bioinformatics can also help in the "Build" process, as automation is becoming more common in the DNA assembly workflow, due to the high popularity of microfluidics devices and liquid-handling robot platforms. In the "Test" phase, computational tools are used to standardize the assays so that comparable results can arise. During all steps of the cycle, large-scale data mining is usually necessary, but it is imperative in the "Learn" phase, where all the knowledge is compiled to give back useful information (Appleton *et al.*, 2017; Konur *et al.*, 2021).

2.4. When the synthetic biology framework meets metabolic engineering

Metabolic engineering emerged at the beginning of the decade of 1990s and has been applied to bioproduction since then. This science addresses the cells as little factories, with thousands of chemical compounds being generated in intrinsically intertwined biochemical reactions (Stephanopoulos, 2012). Thus, the main goal of its studies is to engineer those biochemical pathways from the metabolism of living systems and alter them to direct the metabolic flux toward the production of some specific compound of interest. In this context, microorganisms are turned into microbial cell factories, modified to utilize renewable substrates to produce the desired chemicals, even if they are not native to their metabolism (Cho *et al.*, 2022). Among the approaches belonging to the metabolic engineering structure are the following:

a) Enzyme engineering, which consists in changing enzyme catalytic activity or efficiency by random or site-specific mutagenesis, improving promoter strength or gene copy number, or subcellular colocalization strategies.
b) Pathway engineering, consisting in rewiring the entire carbon flux of the biosynthetic pathway. This can be accomplished by heterologous expression of genes, improving the supply of enzyme substrates and cofactors, reducing negative feedback, downregulating competing pathways, or even developing alternative routes.
c) Cell engineering, which consists in enhancing the stability and productivity of strains to maintain the desired phenotype even when facing harsh conditions. This robustness can be achieved by improving the secretion of

toxic chemicals, adaptive laboratory evolution and chromosome engineering.

With the rise of synthetic biology a few years later, the availability of genetic tools increased, allowing for more audacious projects. Also, the fine-tuning of computational tools opened doors for a deeper exploration of metabolic pathways, and the pursuit of more genes, promoters, terminators, transcription factors and other genetic elements to be applied in such projects. Thus, expanding the possibilities of carbon flux rewiring to new levels of complexity. Besides, all the strategies mentioned above included in the DBTL cycle, since designing a project, building genetic circuits with state-of-the-art methodologies and testing them in biological chassis, can also be applied to metabolic engineering projects (Stephanopoulos, 2012; Gong, Nielsen and Zhou, 2017; Cho *et al.*, 2022; Li *et al.*, 2022).

3. Metabolic engineering of non-conventional yeasts

The model bacterium *Escherichia coli* and the model yeast *Saccharomyces cerevisiae* have been used as microbial cell factories for decades, especially for being well adapted to laboratory conditions and having easy-manipulable genomes. Recently, however, the scientific community has realized a lack of robustness in these model organisms, unable to withstand industrial conditions from a wide range of biotechnological applications. Luckily, several other microorganisms have been studied to be used as alternatives to the conventional ones, which are usually called non-conventional organisms (Nora, Westmann, *et al.*, 2019). These can be organisms that present favorable traits such as higher growth rates, complex post-translational modifications, resistance to fluctuating environments, or extreme temperatures, and that can grow in more complex substrates like lignocellulosic hydrolysates (Vogl, Hartner and Glieder, 2013; James M Wagner and Alper, 2016; Park, Nicaud and Ledesma-Amaro, 2018). Some yeast species are good examples of non-conventional organisms with one or more of these traits: *Yarrowia lipolytica, Kluyveromyces sp., Pichia pastoris* and *Rhodosporidium toruloides*. Diverse molecular toolboxes combining synthetic biology and metabolic engineering were developed over the years to engineer these yeast strains for exploring capabilities (Nora, Westmann, *et al.*, 2019).

For instance, *Y. lipolytica* is an oleaginous yeast capable of assimilating alkanes and producing fatty acid-derived products, such as omega-3 eicosapentaenoic acid and lycopene. Other compounds have also been shown to be produced like citric acid, erythritol (James M Wagner and Alper, 2016; Darvishi *et al.*, 2018), and several different terpenes, such as limonene, pinene, squalene, carotenoids, among others (Li *et al.*, 2022;

Zhang *et al.*, 2022). This non-model yeast has had many elements characterized, such as promoters, terminators, and selection markers (James M. Wagner and Alper, 2016; Darvishi *et al.*, 2018). Tools built for engineering this yeast include: *loxP*/Cre recombination system (de Pourcq *et al.*, 2012) and integrative vectors such as pYL15 (Bredeweg *et al.*, 2017). Recently, much effort has been dedicated to creating CRISPR technologies for genome editing of this oleaginous yeast (Schwartz *et al.*, 2018; Shi *et al.*, 2018; Yang, Edwards and Xu, 2020). Additionally, a dedicated toolkit of Golden Gate Assembly parts was also developed for this organism (Celińska *et al.*, 2017; Larroude *et al.*, 2019).

Regarding the genus *Kluyveromyces,* several species have already demonstrated biotechnological importance. One of them is *Kluyveromyces lactis,* mainly used to produce enzymes for dairy products, such as the milk-digesting enzyme lactase and the milk-clotting enzyme chymosin (Van Ooyen *et al.*, 2006; Spohner *et al.*, 2016). Another example is *Kluyveromyces marxianus,* a thermo-tolerant yeast used to produce biofuels in high-temperature conditions (Yang *et al.*, 2015; Nambu-Nishida *et al.*, 2017). Development of tools for these two non-conventional yeasts include loxP/Cre methodologies (Ribeiro *et al.*, 2007), integrative vectors (Colussi and Taron, 2005; Lambertz *et al.*, 2016; Nambu-Nishida *et al.*, 2017), and CRISPR technologies that have been constantly improved and implemented (Löbs *et al.*, 2017; Hwang, Lee and Lee, 2018; M. Li *et al.*, 2021; Rajkumar and Morrissey, 2022). *K. marxianus* also had a toolkit of Golden Gate Assembly parts recently developed (Rajkumar and Morrissey, 2020).

Moreover, *P. pastoris* is frequently used for heterologous protein production – it is the most commonly used eukaryote for single protein expression (Vogl, Gebbie, Robin

W. Palfreyman, *et al.*, 2018). This species has been widely used for its capacity to make fully humanized glycosylated proteins (Vogl, Hartner and Glieder, 2013). Because of that, this yeast now has a well-established promoter toolset (Vogl and Glieder, 2013; James M. Wagner and Alper, 2016). Also, with high-quality genome sequences are available for strain engineering (De Schutter *et al.*, 2009; Küberl *et al.*, 2011), most of the tools available are based on genomic integration (Vogl, Gebbie, Robin W Palfreyman, *et al.*, 2018). Some strategies used for genetic engineering of this yeast were *LoxP/Cre* technology (Pan *et al.*, 2011), FLP recombinase system (Näätsaari *et al.*, 2012), a multi-part toolkit (Obst, Lu and Sieber, 2017), and a dedicated Golden Gate platform as well (Prielhofer *et al.*, 2017).

Yet, *R. toruloides* is a Basidiomycete fungus capable of accumulating a high amount of lipids and carotenoids in its cells and it is, therefore, called oleaginous and carotenogenic yeast. Differentiating itself from the previously mentioned non-conventional yeasts, *R. toruloides* stands out for its innate ability to consume extremely complex substrates and survive overly toxic environments. Even though molecular tools have been developed for the other non-conventional yeasts for a more extended time, there are currently several tools being established for *R. toruloides* as well. The following section will be dedicated to presenting the favorable traits that make this emergent host of such importance to biotechnology industries and how it can be used to produce high-value-added chemicals.

4. *R. toruloides* as a cell factory for chemical production

The oleaginous yeast *R. toruloides* is emerging as a promising microbial cell factory. One of the primary reasons for using this yeast as a host for application in biorefineries is its ability to utilize a wide range of different substrates, including agroindustrial residues. Regarding lignocellulosic biomass, *R. toruloides* can consume a significant proportion of it, including the co-consumption of five-carbon sugars, like xylose (C5), and six-carbon sugars, like glucose (C6). This organism can also consume some of the compounds derived from lignin hydrolysates, such as p-coumaric and ferulic acid (Yaegashi et al., 2017). This oleaginous yeast is, therefore, more robust when compared to the previously mentioned industrial hosts. For instance, *R. toruloides* naturally produces and accumulates lipids up to 60% of its dry mass (Park et al., 2018), which is more than the other oleaginous yeast *Y. lipolytica*. It can also grow on a broader range of substrates than all other strains. Some examples of substrates successfully applied for the growth of *R. toruloides* are sugarcane juice and bagasse, crude glycerol, wheat straw, cassava starch, among others (Bonturi *et al.*, 2017; Park, Nicaud and Ledesma-Amaro, 2018; Lopes, 2020).

R. toruloides can produce large amounts of products valuable to industry. For instance, this carotenogenic yeast can naturally produce very high yields of carotenoids, the precursors of vitamin A, and can be used as colorants and antioxidants (Park, Nicaud and Ledesma-Amaro, 2018). Yet, by applying synthetic biology combined with metabolic engineering, the possibilities of bioproduction increased at an even higher rate. Some of the bioproducts that were produced when this yeast was grown in different kinds of agroindustrial residues can be seen in **Figure 2**.

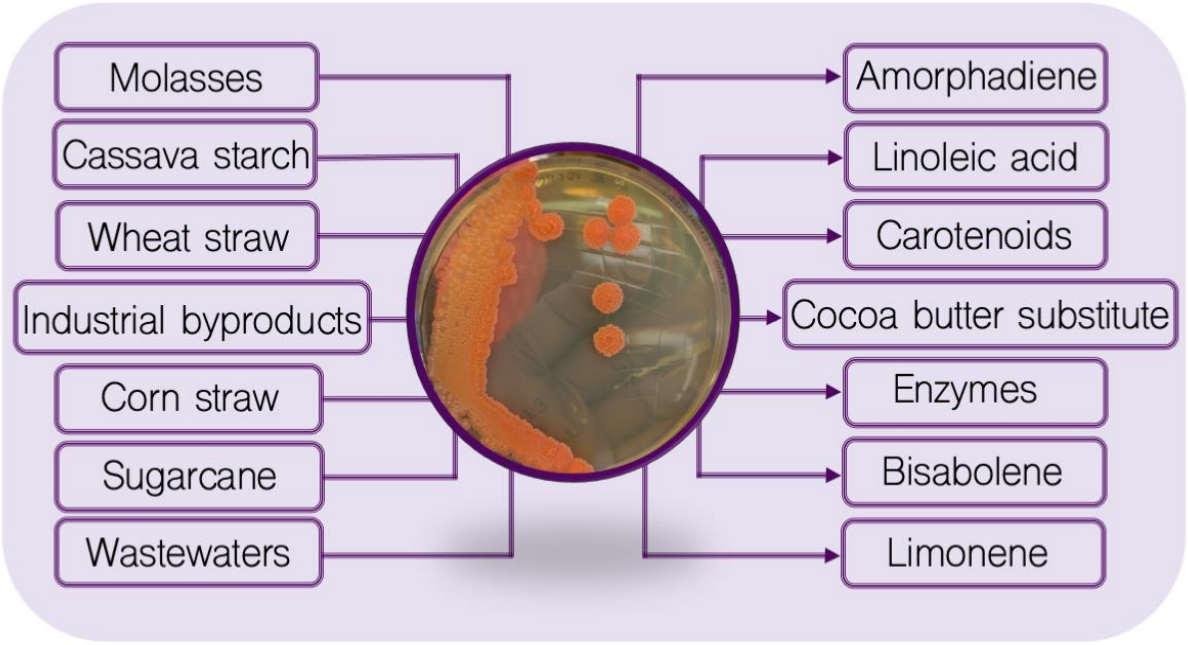


Figure 2. Bioproduction in *R. toruloides* using a wide range of carbon sources, including agro-industrial residues.

In order to successfully implement synthetic biology for the application of metabolic engineering strategies in this yeast, the development of specific tools is required. In this regard, specific promoters have already been characterized, including some constitutive (Liu *et al.*, 2013; Wang *et al.*, 2016), inducible (Alexander M.B. Johns, Love and Aves, 2016), and some strong intronic promoters (Liu *et al.*, 2016a). More recently, a toolset of more than 20 promoters was characterized by Nora et al., (2019), enclosing some strong bidirectional ones (Nora, Wehrs, Kim, Skerker, *et al.*, 2019). Although several research groups are looking for it, episomal DNA was never reported for this yeast, so plasmid DNA cannot be used for DNA manipulation. Instead, the only way of altering the DNA is to have homologous recombination and integrate the constructs in its genome. Because of that, one of the most common transformation methods used up until now is

Agrobacterium tumefaciens-mediated transformation (ATMT) for single or multiple integrations (Liu *et al.*, 2013; Lin *et al.*, 2014). Additionally, there have been reports of transformation by electroporation (Liu *et al.*, 2017), but, in the last few years, most researchers have been focusing on the lithium acetate transformation method (Tsai *et al.*, 2017; Nora, Wehrs, Kim, Skerker, *et al.*, 2019; Bonturi, Pinheiro, Monteiro de Oliveira, *et al.*, 2022). The selection markers most commonly used for this species are antimicrobials such as hygromycin B, nourseothricin, bleomycin and G418 (Lin *et al.*, 2014; Alexander M.B. Johns, Love and Aves, 2016). Additionally, CRISPR-based genome editing tools (Jiao *et al.*, 2019; Otoupal *et al.*, 2019; Schultz, Cao and Zhao, 2019) and RNA interference tools (Liu *et al.*, 2019) have been developed for more efficient DNA manipulation. The latest development in genetic engineering tools was a Golden Gate platform designed especially for this organism (Bonturi, Pinheiro, Monteiro de Oliveira, *et al.*, 2022). Also, several genome-scale models (Dinh *et al.*, 2019; Tiukova *et al.*, 2019; Kim *et al.*, 2021) and functional genomics (Coradetti *et al.*, 2018) is advancing our understanding and enabling us to perform metabolic engineering of *R. toruloides* for production of value-added compounds (Wen *et al.*, 2020; Cao *et al.*, 2022).

Several metabolic engineering studies for bioproduct formation in *R. toruloides* have focused on lipid accumulation (Zhang *et al.*, 2016; Wang *et al.*, 2018). However, the considerable accumulation of lipids suggests suitability for applications in any product derived from acetyl-CoA. These applications include fatty alcohols, methyl ketones, polyketides and isoprenoids, translating into applications ranging from fuels and specialized chemicals to complex drugs (Yuan and Ching, 2016; Campbell, Xia and Nielsen, 2017; Yaegashi, Kirby, Ito, Sun, Dutta, Mirsiaghi, Eric R Sundstrom, *et al.*, 2017;

Park, Nicaud and Ledesma-Amaro, 2018). In that sense, one of the main classes of specialized chemicals that have their production expanding over the last few years are terpenes.

5. Terpene production in *R. toruloides*

Terpenes are a substantial class of natural products composed of various isoprenoid chains. They are widely used today for various purposes, including fragrances, food additives, pharmaceuticals and biofuels (Monteiro and Veloso, 2004; Paduch *et al.*, 2007; Geiselman *et al.*, 2020; Y. Liu *et al.*, 2021). They are commonly extracted from plants, but the high cost of these plants and the constant competition with food production make the microbiological production of terpenes, which is sustainable and scalable, more interesting for the industry. The advantages include the high productivity and an abundance of low-cost biomass to be used as raw material (Bastos, 2007; Pacheco, 2011; Sorek *et al.*, 2014; Soccol *et al.*, 2017; Martins-Santana *et al.*, 2018). Examples of some terpenes already being produced on a large scale by microorganisms range from lubricants to malaria drugs (Tanasova and Sturla, 2012; Kang and Lee, 2015; Benjamin *et al.*, 2016). For example, ß-farnesene is a sustainable alternative that can be used as a precursor of adhesives, fragrances, lubricants, surfactants, resins, and emulsifiers, among other applications. Biotech company Amyris® produces this compound using genetically modified *S. cerevisiae*, which sells for US$ 100 per kilogram in the United States (http://www.farnesene.net). Another example is alpha-humulene, which has recently gained attention because it has demonstrated anti-inflammatory properties as effective as conventional drugs (Fernandes *et al.*, 2007), in addition to having been reported to have antimicrobial and antitumor activity (Pichette *et al.*, 2006; Legault and Pichette, 2007). Another critical commercial terpene is artemisinin, a recognized drug against malaria. This compound used to be extracted from the plant *Artemisia annua* and, thanks to the advent of synthetic biology, can now be produced by yeasts (Peplow, 2016).

Terpenes are synthesized from two isomeric C5 building blocks: isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP). They are divided into classes based on the number of isoprene units in the molecule. The prefix indicates the number of isoprene pairs assembled in the molecule: monoterpenes (C10), diterpenes (C20), sesquiterpenes (C15), sesterterpenes (C25), triterpenes (C30), and tetraterpenes (C40). In microorganisms, IPP and DMAPP can be produced using two biosynthetic pathways. In Bacteria, these universal precursors are produced in the 2-C-methyl- D-erythritol-4-phosphate (MEP) pathway. Yet, for Eukaryotes and Archaea, the mevalonate (MVA) pathway is the main one (**Figure 3**). Thus, terpenes production starts from the precursor acetyl-CoA, which comes from glycolysis and PDH or citrate pathways. Two molecules of acetyl-CoA are condensed by thiolase (ERG10) to form acetoacetyl-CoA. A second condensation by hydroxymethylgluataryl-CoA synthase (ERG13) forms 3-hydroxy-3-methyl-glutaryl-CoA and the reduction of this molecule by HMG-CoA reductase (HMG1) results in mevalonate. Three reactions convert mevalonate into IPP that is then isomerized to DMAPP. IPP and DMAPP are then condensed in different ratios to form the precursors GPP, FPP, and GGPP. GPP, FPP, and GGPP are synthesized by geranyl pyrophosphate synthase (GPPS), farnesyl pyrophosphate synthase (FPPS), and geranylgeranyl pyrophosphate synthase (GGPPS), respectively. These precursors have different chain lengths and will be the basis for forming the terpenes respective to their own chain lengths, when under the activity of the corresponding terpene synthase, as seen in **Figure 3**.

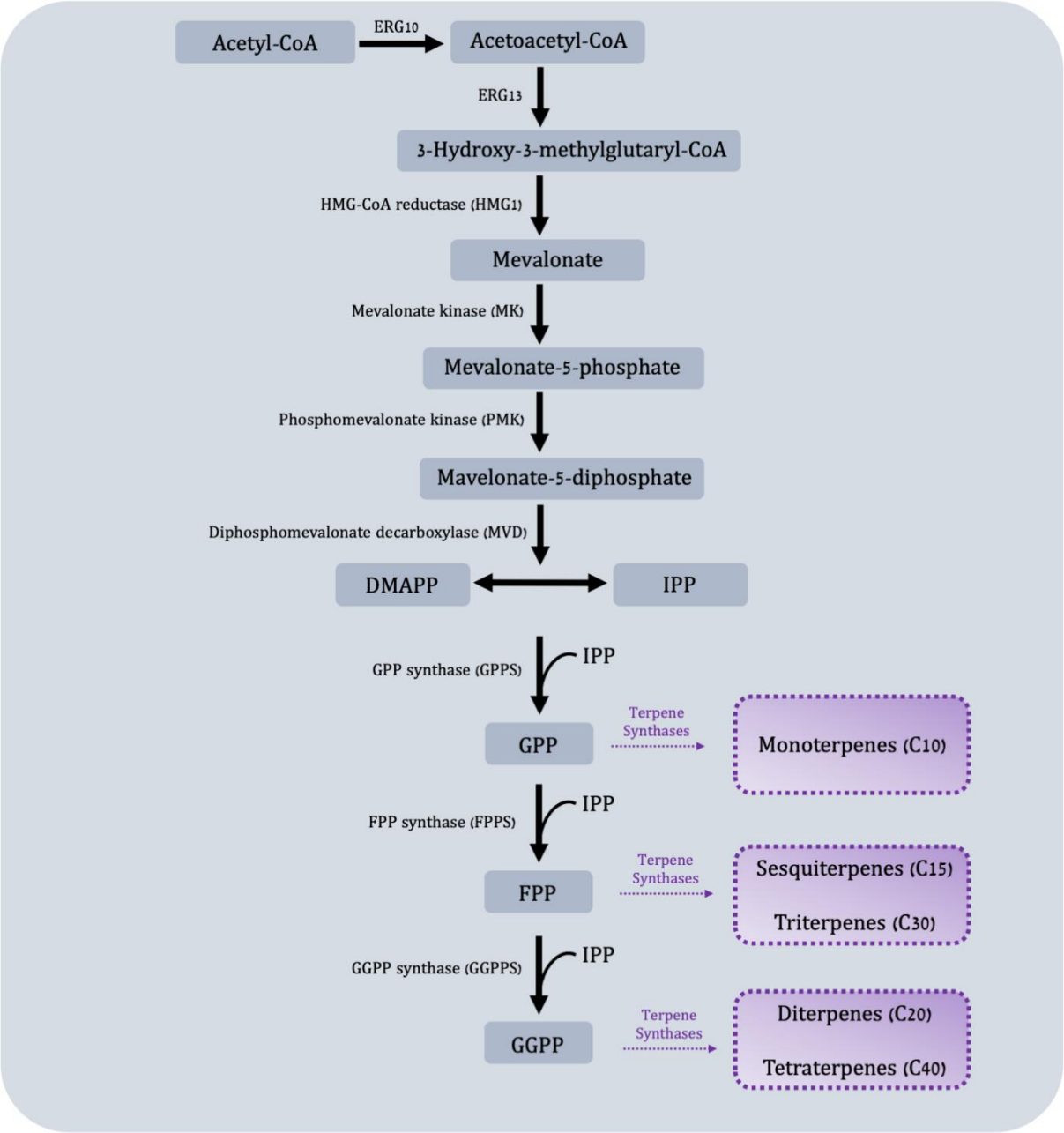


Figure 3. From Acetyl-CoA to terpene precursors through the mevalonate pathway.

Regarding producing terpenes, *R. toruloides* have some advantages over other yeasts. The first reason is that the oleaginous yeast is capable of accumulating acetyl-CoA, the first precursor of the mevalonate pathway. Secondly, this orange yeast is already a good source of carotenoids, which are tetraterpenes. Additionally, there are examples of studies applying metabolic engineering to optimize terpenes production that has successfully produced several of the classes of terpenes (Martins-Santana *et al.*, 2018; Park, Nicaud and Ledesma-Amaro, 2018). This shows the potential of this organism to serve as a host in the production of these chemicals of industrial and pharmaceutical interest.

For instance, Yaegeshi *et al.* (2017) demonstrated the high-yield production of amorphadiene, the precursor of artemisinin, and bisabolene, a precursor of biodiesel, in *R. toruloides*, with no other modifications besides the expression of the exogenous terpene synthase enzyme. Kirby *et al.* (2021) later engineered this same strain obtaining even higher yields of bisabolene production by adding more copies of the gene to the genome. This work also presented an increased production of 1,8-cineole, one of the monoterpenes that can be used as biofuel. They achieved this high production not only by adding multiple copies of the heterologous monoterpene synthase, but also by adding a less efficient FPPS enzyme, which reduces the flux of carbon to competing pathways. Zhuang *et al.* (2019) expressed 16 genes encoding different monoterpene synthases. From those, eight were successfully expressed and the strains were able to produce monoterpenes, but still achieving low yields. Yet, in 2021, Liu *et al.* (2021) created *R. toruloides* strains producing the monoterpene limonene, which can be used as biofuel and as an additive in foods, cosmetics, and cleaning products. Limonene production was

achieved by heterologous expression of several enzymes from the mevalonate pathway, while down-regulating carotenoid production, which competes for the enzyme substrates. A compilation of important terpenes that have been shown to be produced by *R. toruloides* can be seen in **Table 1**. Substantial improvement has been made in terpene production in this yeast in the last few years, but it is still modest compared to other conventional and non-conventional yeasts. This was mainly due to the need for more genomic information and gene manipulation technologies - which is slowly being fulfilled, as mentioned previously. For this reason, the rate of success for heterologous expression of terpene synthases in this organism should increase in the upcoming years.

Table 1. Terpene production in *R. toruloides*

Terpene class	Terpene name	Titer	Condition	Scale	Reference
Monoterpene	1,8-cineole	160 mg/L	Rich media	Flasks	Kirby *et al*. (2021)
Monoterpene	1,8-cineole	1.2 g/L	Rich media	Flasks	Kirby *et al*. (2021)
Monoterpene	1,8-cineole	1.4 g/L	Hydrolysates	Flasks	Kirby *et al*. (2021)
Sesquiterpene	α-bisabolene	2.2 g/L	Hydrolysates	Bioreactor	Kirby *et al*. (2021)
Monoterpene	Limonene	117.8 mg/L	Rich media	Flasks	Liu *et al.* (2021)
Diterpene	ent-kaurene	1.4 g/L	Hydrolysates	Bioreactor	Geiselman *et al.* (2020)
Monoterpene	1,8-Cineole	34.6 mg/L	Hydrolysates	Bioreactor	Zhuang *et al.* (2019)
Monoterpene	Ocimene	<5mg/L	Rich media	Flasks	Zhuang *et al*. (2019)
Monoterpene	Limonene	<1mg/L	Rich media	Flasks	Zhuang *et al.* (2019)
Monoterpene	Pinene	<5mg/L	Rich media	Flasks	Zhuang *et al.* (2019)
Monoterpene	Sabinene	<5mg/L	Rich media	Flasks	Zhuang *et al.* (2019)
Monoterpene	Carene	<5mg/L	Rich media	Flasks	Zhuang *et al.* (2019)
Sesquiterpene	Amorphadiene	36 mg/L	Rich media	Flasks	Yaegashi et al., 2017
Sesquiterpene	α-bisabolene	680 mg/L	Hydrolysates	Bioreactor	Yaegashi et al., 2017

The inevitability of climate change and global warming intensified the search for more sustainable alternatives of chemical production. In this regard, *R. toruloides* is a significant organism that can be exploited as a microbial cell factory for renewable production of those valuable chemicals. Among them, we can cite terpenes, which can be used for many applications, including biofuels. For that reason, it is extremely important to understand *R. toruloides* physiology and genomic and transcriptomic behaviors and to use that knowledge to further develop new synthetic biology tools for this organism. Moreover, the toolbox under continuous optimization can be applied for metabolic engineering of this oleaginous yeast, to create even more robust strains that can grow on a wide range of substrates, including agro-industrial residues. All these traits can be combined to produce green chemicals, reducing the human impact on Earth and pursuing the circular bio-based economy principles.

II. OBJECTIVES

General Objective

To use synthetic biology-based approaches to understand transcriptional responses and to design and develop tools for improving production of valuable chemicals in *R. toruloides*.

Specific Objectives:

A) To perform RNA sequencing experiments in order to acquire information about transcritional behaviors of *R. toruloides* gowing in conditions simulating industrial ones, and to develop a new bioinformatics pipeline for discovery of *cis*-regulatory elements using this transcritomic data as input.

B) To apply metabolic engineering methods in order to create *R. toruloides* strains capable of producing three monoterpenes of commercial interest: pinene, linalool and geraniol.

III. RESULTS

CHAPTER ONE

Mining novel cis-regulatory elements from the emergent host *Rhodosporidium toruloides* using transcriptomic data

This chapter is now published in Frontiers in Microbiology under the DOI: https://doi.org/10.3389/fmicb.2022.1069443.

ABSTRACT

The demand for robust microbial cell factories that can produce valuable biomaterials while being resistant to stresses imposed by current bioprocesses is rapidly growing. *R. toruloides* is an emerging host that presents desirable features for bioproduction, since it can grow in a wide range of substrates and tolerate a variety of toxic compounds. In order to explore *R. toruloides* suitability for application as a cell factory in biorefineries, we sought to understand the transcriptional responses of this yeast when growing under experimental settings that simulated those used in biofuels-related industries. Thus, we performed RNA sequencing of the oleaginous, carotenogenic yeast in different contexts. The first ones were stress-related: two conditions of high temperature (37 ˚C and 42 ˚C) and two ethanol concentrations (2% and 4%), while the other was using the inexpensive and abundant sugarcane juice as substrate. Using transcriptomic data, differential expression and functional analysis were implemented to select differentially expressed genes and enriched pathways from each set-up. A reproducible bioinformatics workflow was developed for mining new regulatory elements. We then predicted, for the first time in this yeast, binding motifs for several transcription factors, including HAC1, ARG80, RPN4, ADR1, and DAL81. Most of the putative transcription factors uncovered here were involved in stress responses and found in the yeast genome. Our method for motif discovery provides a new realm of possibilities in the study of gene regulatory networks, not only for the emerging host *R. toruloides*, but for other organisms of biotechnological importance.

Keywords: transcription factor, motifs, RNA sequencing, sugarcane juice, industrial stress, cell factory, biorefineries.

Abbreviations:

DEG – differentially expressed genes

GO – Gene Ontology

KEGG – Kyoto Encyclopedia of Genes and Genomes

KOG – EuKaryotic Orthologous Groups

PCC – Pearson's correlation coefficient

SCJ – sugarcane juice

TF – transcription factor

TFBS – transcription factor binding site

1. Background

Microbial cell factories are in ever-growing demand due to the pursuit of more sustainable products that support the concepts of green chemistry and biorefineries (Cherubini, 2010; Martins *et al.*, 2021). The benefits go beyond obtaining valuable products from low-cost raw materials, but include the independence from oil, a finite and extremely polluting material (Borodina and Nielsen, 2014). Innovations on obtaining products from renewable substrates are based on the genetic improvement and metabolic engineering of microorganisms that are able to degrade and ferment different biomasses (Martins-Santana *et al.*, 2018). The main challenges the currently available cell factories face is not only the capability to break complex compounds into simpler chains, but also to be able to tolerate adverse conditions resulting from bioprocesses (Fletcher *et al.*, 2017).

Rhodosporidium toruloides is emerging as a potential, robust host for the completion of both tasks. This oleaginous, carotenogenic yeast can flourish on a wide range of substrates, such as sugarcane juice and bagasse, crude glycerol, wheat straw, cassava starch, among others (Bonturi *et al.*, 2017; Park, Nicaud and Ledesma-Amaro, 2018; Lopes, 2020). It can also store high amounts of lipids – up to 60% of its total cell weight – and tolerate compounds commonly toxic for other yeast species (Saini *et al.*, 2020). It performs better than the widely used baker's yeast, *Saccharomyces cerevisiae*, for example, as it can degrade C5 sugars and lignin-derived compounds, which the latter cannot (Nora, Westmann, *et al.*, 2019).The advantages go over other lipid-producing strains as well, such as *Yarrowia lipolytica*, since it can grow in more complex carbon sources and presents higher tolerance to several inhibitors like 5-hydroxy methyl ester,

furfural, acetic acid and vanillin (Saini *et al.*, 2020). *R. toruloides* has already been shown to produce several different bioproducts in a variety of conditions. Examples of these bioproducts include enzymes, carotenoids and lipid-based compounds: linoleic acid, oleic acid and cocoa butter substitute (Park, Nicaud and Ledesma-Amaro, 2018). Recently, terpenes that can be used as precursors of biodiesel and jet fuel alternatives were produced in pilot scale, derived from corn stover hydrolysates (Yaegashi, Kirby, Ito, Sun, Dutta, Mirsiaghi, Eric R Sundstrom, *et al.*, 2017; Kirby *et al.*, 2021). Hence, all these distinctive characteristics point to *R. toruloides* as a great candidate to improve the productivity of biorefineries.

There is great interest in using unconventional, more robust yeasts in detriment to the use of *S. cerevisiae* for ethanol production in bioreactors (Wehrs *et al.*, 2019; Wu *et al.*, 2020). Ethanol tolerance and thermotolerance, for instance, are desirable features. First, ethanol accumulation during fermentation can become a significant stress factor during the process (Stanley *et al.*, 2010). Secondly, higher fermentation temperatures can lower the cost of cooling bioreactors during the process and reduce bacterial contamination, decreasing the need for antibiotics (Abdel-Banat *et al.*, 2010; Huang *et al.*, 2018). Thus, in order to explore if *R. toruloides* could tolerate and thrive in these stressful environments, we decided to investigate the transcriptional response of this yeast growing in high temperatures and in the presence of ethanol. Furthermore, Soccol *et al.* (2017), developed a process of producing biofuels using *Rhodosporidium* grown in a simple media containing only sugarcane juice and urea. Sugarcane juice is an abundant and low-cost substrate, containing up to 15% of fermentable sugars. Based on their work,

we additionally sought to understand the changes in transcription of *R. toruloides* grown in media containing sugarcane juice and urea (Soccol *et al.*, 2017).

Transcription factors (TF) are proteins that bind to specific DNA sequences modulating the rate of transcription of DNA to mRNA. These specific DNA sequences are collectively called transcription factor binding sites (TFBS), or binding motifs, and are present in the *cis*-regulatory region of DNA surrounding the transcription starting sites. The activity of TFs can act directly or indirectly to transform physiological and environmental signals into patterns of gene expression. Thus, identifying both the proteins and the position of their binding motifs is essential for understanding gene regulation in the cell and allowing better transcription control, which can serve as a basis for new bioengineering applications (Martins-Santana et al., 2018). However, finding new regulatory sequence motifs is not an easy task, especially in eukaryotic organisms. One of the reasons is that primary nucleotide sequence is not the only characteristic that specifies a target. Additional factors, such as genomic context, DNA binding, DNA modification and DNA shape can change nucleotide preferences (Inukai, Kock and Bulyk, 2017). Besides, it is impossible to test all the environmental conditions that a natural regulatory network is able to respond to. Nevertheless, some tools were developed aiming to predict regulatory elements in specific datasets, helping to overcome limitations in motif discovery.

Here, we developed a new pipeline for discovery of *cis*-regulatory elements integrating tools for motif prediction and applied it to the emergent microbial cell factory *R. toruloides*, which can later be expanded to other biotechnologically relevant organisms. In order to do that, we cultivated *R. toruloides* in the settings described above and

extracted RNA samples from early time points of growth. Then, we performed RNA sequencing, functional analysis of the transcriptional responses and discovery of new motifs from genes of the main pathways found to be enriched in each specific condition. In this sense, we are providing new insights regarding the regulation of *R. toruloides* genes in a transcriptomic approach and describing, for the first time, putative TFBS that can be used for transcriptional control in this yeast, offering the basis for building new genetic tools for this promising host.

2. Material and Methods

2.1. Strains, media and growth conditions

For RNA sequencing experiments using *R. toruloides* grown in a medium with sugarcane juice (SCJ), cultures were first grown in LB medium (1% yeast extract, 1% tryptone and 0.5% NaCl) for 24 hours. LB medium was used in the inoculum so that there was no other sugar influencing the yeast metabolism. The inoculum was centrifuged and transferred to media containing sugarcane juice. Considering the concentration of 120g of sucrose per liter of juice, we standardized the final concentration of the media as 4% sucrose. In addition, 1% urea was added to be used as a nitrogen source. The filtered sugarcane juice was pasteurized at 65˚C for 30 minutes before being added to the autoclaved medium. SCJ cultures were grown for 8 hours and then collected for RNA extraction. Aliquots of the inoculum in LB grown for 24 hours were also collected to serve as internal control and henceforth will be called time 0 h. The 8-hour growth time was

selected for this condition because the purpose was to analyze the genes as soon as there was expression of the genes for adaptation to the new medium.

The experiments for the stress conditions were carried out as follows: the inoculum was grown in YPD medium (1% yeast extract, 2% peptone and 2% glucose) at 200 rpm at 30 ˚C for 24 hours. These inoculants were centrifuged and transferred to the respective media, containing YPD and one of the respective stress conditions: 2% ethanol and 4% ethanol, which were grown in a shaker at 30 ˚C, and the high temperature conditions that were grown at 37 ˚C and 42 ˚C. The control condition was determined as the inoculum grown in YPD for 24 hours. After being transferred to the respective media, the cultures were grown for another 16 hours and collected for RNA extraction and sequencing. All inoculations were performed in a 1:10 ratio, and all experiments were performed in triplicates. The strain *R. toruloides* IFO0880 was used in all experiments (Lin *et al.*, 2014; Coradetti *et al.*, 2018).

2.2. RNA extraction and sequencing

Cell lysis was performed by pelleting yeast cells from the culture by centrifugation and resuspending it with Trizol® (Thermo Fisher Scientific, Vilnius, Lithuania). Samples were transferred to tubes containing zirconium beads and lysed using a cell homogenizer. RNA extraction using Trizol® (Thermo Fisher Scientific, Vilnius, Lithuania) was performed following the manufacturer's instructions. The resulting RNA samples were immediately purified using the Qiagen® Rneasy mini kit. Samples were submitted to analysis by Agilent Bioanalyzer RNA 6000 Nano kit for RNA integrity check, aiming for a RIN number higher than 8. After being extracted and purified, all RNA samples were kept in a -80 ˚C freezer.

The samples were sent for sequencing at the Genomics Center of the University of São Paulo in Piracicaba, São Paulo, Brazil. The sequencing library was prepared by the facility using Illumina TruSeq Stranded mRNA Sample Prep LT Protocol. RNA sequencing was performed using the HiSeq SBS v4 kit in Illumina HiSeq 2500, with paired reads of 100 bp (2x101). Raw data is available as Sequencing Read Archives (SRA) on the NCBI website under accession number PRJNA883675.

2.3. RNA-seq differential gene expression analysis

Read quality check was performed using FastQC (Andrews, 2018) and the trimming was made using the Trimmomatic software (Bolger, Lohse and Usadel, 2014). Henceforth, we used only the reads assigned by Trimmomatic as paired. Gene expression was quantified with kallisto (Bray *et al.*, 2016) using for counting the *R. toruloides* IFO0880 JGI reference transcriptome (Coradetti *et al.*, 2018) available at https://mycocosm.jgi.doe.gov/Rhoto_IFO0880_4/Rhoto_IFO0880_4.home.html. The differential expression analysis was done using DESeq2 (Love, Huber and Anders, 2014) in the R platform (R Core Team, 2019). Genes whose adjusted *p*-values (*adjpval*) were lower than 0.05 and whose log2-fold change values were lower than −1 or higher than 1 were selected as differentially expressed genes (DEGs).

2.4. Functional and pathway analysis

In order to analyze the DEGs function, we used the reference gene function annotation file created by KOG (also provided by JGI, as above). For pathway enrichment analysis we used the GAGE R package (Luo *et al.*, 2009). As input in this step, we used the DEGs from each condition with their function annotated by the KEGG Automatic Annotation Server (https://www.genome.jp/tools/kaas/), as this tool provides a compatible file for GAGE analysis. The enriched pathways were filtered by a p-value lower than 0.05. Bubble maps were plotted using R package ggplot2 (Wickham, 2016). Pathview R package (Luo and Brouwer, 2013) was used to visualize enriched pathways whenever needed.

2.5. Putative *cis*-regulatory elements discovery.

For each condition, we selected the DEGs that belong to the KOG classes. We extracted the promoter sequence of these genes on the reference database and applied the HOMER software (Heinz *et al.*, 2010) in each promoter sequence sets, with the following parameters: -len 10, -stat hypergeo, -olen 3, -strand both, -maxBack 0.3. The background file was the *R. toruloides* IFO0880 JGI reference promoter sequences. We performed a pairwise comparison between all motifs, by applying the universalmotif (Tremblay, 2020) function ‘compare_motifs’ using Pearson Correlation Coefficient (PCC) as method and average mean as scoring strategy. We also performed pairwise alignment using TOMTOM, from MEME suite (Gupta *et al.*, 2007), providing as input our motif file as both target and source, with the following arguments: -min-overlap 0, -evalue, -thresh 10.0 and -dist pearson. Finally, we filtered the results by a p-value lower than 0.05, an e-

value lower than 1 and a PCC higher than 0.8. All motifs from our experimental settings were found using HOMER and were aligned with the ones from JASPAR2018_CORE_fungi_non-redundant database using TOMTOM. Subsequently, they were plotted and analyzed as a graph, using Gephi (Bastian, Heymann and Jacomy, 2009), applying Circle Pack Layout for representation with the variables 'predicted or annotated', experimental condition and node degree as hierarchy.

2.6. Finding TF candidates

For the overrepresented motifs found on JASPAR alignment, we got the amino acid sequence of the corresponding transcription factor protein and then performed a blast on JGI platform, to search for similar proteins on the *R. toruloides* reference assembled genome (provided by JGI, as above). Amino acid sequences and GO functions of each predictive TF were extracted from UniProt (http://uniprot.org).

2.7. Bioinformatics Workflow

All analyses performed on the R environment were organized in an easy-to-use Jupyter Notebook, the customized code and instructions to install and run the workflow are publicly available at https://github.com/computational-chemical-biology/cis_reg.

3. Results

3.1. A new pipeline for discovery of regulatory elements from transcriptomic data

Figure 1 summarizes the workflow developed in this work. The experimental design consisted of growing cells and extracting their RNA as described in the Methods section. Then, the RNA sequencing data was used as input for the motif discovery pipeline. First, differential expression analysis was performed in order to select the Differentially Expressed Genes (DEGs), and functional analysis was executed to find enriched pathways for each condition. We then extracted the promoter sequences from the genes of the specific enriched pathways and applied the HOMER software for motif prediction. **Table 1** contains the number of promoter sequences and motifs found by HOMER for each case. We performed an internal pairwise comparison between all motifs, by applying both the universalmotif package and TOMTOM tool. Subsequently, we performed an external comparison against the Jaspar database using TOMTOM. The entire workflow will be described in detail throughout the Results section.

3.2. Differential expression analysis shows a significant transcriptional change in sugarcane and high temperature conditions

The RNA sequencing data were analyzed through the RNA-seq differential analysis using DESeq2. For all analyses, there were two different condition settings: the first one was the sugarcane juice condition (SCJ) and the second one was the stress

conditions (comprising 37 and 42 ˚C and ethanol 2% and 4%). The differential expression of *R. toruloides* genes grown on the SCJ condition was measured relative to its own control, which was the inoculum grown in LB for 24 hours (also called time 0h). Yet, the differential expression of genes from *R. toruloides* grown in stress conditions were measured using the inoculum grown in YPD for 24 hours. Each experimental setting was always analyzed separately since they had different internal controls. The principal component analysis (PCA) was performed to check the quality of the sequencing samples, where we could observe the consistency of our sample replicates in a summarized representation. As shown in the PCA graphs in **Supplementary Figure S1A**, where the two first components explain 84% of the variance, the replicates of each condition are close to each other and distant from their respective controls, which indicates that the gene expression differences are consistent inside each condition, allowing the comparison between them. Nevertheless, in **Supplementary Figure S1B**, we can see that the replicates of the culture grown at 37 ˚C are very close to the control condition. While the 42 ˚C is separated by the first principal component, explaining alone 50.1% of the variance.

To further investigate the quality of the RNA-seq experiments, the distance between samples, based on the transcript expressions, was measured (**Supplementary Figure S2**). In the heatmap analysis, Euclidean distance was applied to the expression pattern of each sample, where darker blue means more proximity and lighter blue means greater distance between samples. We can see the same trend in which the replicates of each condition are grouped, which confirms the PCA analysis. Similar to PCA, in **Supplementary Figure S2A**, we can see that the ethanol 2% and ethanol 4% conditions

were not distant from each other. This shows that the changes in transcriptional behavior in *R. toruloides* were not divergent between the two conditions, albeit they are both distant from the control. In **Supplementary Figure S2B** it is possible to see a similar trend to PCA, where the SCJ condition is distant from its internal control.

The Venn Diagram depicted in **Figure 2** summarizes the DEGs found for *R. toruloides* grown in each experimental condition. The conditions in which more genes presented transcriptional change were the SCJ and the temperature 42 ˚C. More than three thousand DEGs were found for those experimental settings, while there were around one thousand DEGs for each of the other conditions: 37 ˚C and ethanol 2% and 4%. These results show that the latter conditions, although causing a significant change in the transcriptional behavior of *R. toruloides*, did not cause as much impact as the change to a rich substrate caused by the addition of sugarcane in the media nor the increase in temperature by growing at 42 ˚C. *R. toruloides* grown in SCJ had a greater number of down-regulated genes versus up-regulated genes, while growing in 42 ˚C resulted in the opposite: more genes were up-regulated (**Supplementary Figure S3**). The outermost numbers in the Venn Diagram represent the amount of genes that were exclusively differentially expressed for each condition. The exclusive DEGs can lead to the creation of a new inducible promoter library, given that the promoters will likely be induced by the exact same condition in which their genes were differentially expressed. Although still under development, the current promoter toolbox for *R. toruloides* is limited (Alexander M B Johns, Love and Aves, 2016; Liu *et al.*, 2016b; Wang *et al.*, 2016; Nora, Wehrs, Kim, Cheng, *et al.*, 2019). Our dataset can be useful for the development of inducible promoters in the future.

3.3. Functional analyses uncover KEGG pathway enrichment in each condition

To analyze the function of the DEGs, we performed a pathway enrichment analysis. In order to do that, we first had the transcript function annotated by KEGG Orthology. Functional annotation by KEGG was the input required for the R package used for the analysis, called GAGE. Both the annotation and GAGE analysis are described in the methods section. The resulting pathways were filtered to a p-value of less than 0.05. In **Figure 3**, the biochemical pathways of *R. toruloides* that are enriched when using SCJ as a substrate can be seen. The main up-regulated pathways enriched in this condition are those related to cell growth and metabolism, such DNA metabolism and replication, meiosis and amino acid biosynthesis. Pathways related to the biosynthesis of amino acids, ribosomes and cell cycle showed the greatest numbers of up-regulated genes.

For the 42 ˚C condition, although there was a higher number of up-regulated genes (**Supplementary Figure S3**), most of the enriched pathways were found to be down-regulated **(Figure 4)**. Included on that list are several metabolism pathways, showing the exact opposite of what happens in SCJ – metabolism genes are being turned off. In contrast, the condition of 37 ˚C resulted in a reduced number of differentially expressed transcripts when compared to the control grown at 30 ˚C (**Supplementary Figure S4**). Yet, in **Supplementary Figure S5** the routes found to be enriched by GAGE by the presence of ethanol in the media can be seen. Most of the pathways are metabolism-related ones that are being down-regulated in both conditions.

3.4. Identification of putative *cis*-regulatory elements and transcription factor candidates

Subsequently, we performed a new functional analysis, this time using the function annotation file of the reference gene created by KOG (provided by the JGI website as described in Methods). The new functional annotation using the KOG database was chosen due to a better correspondence with our transcript IDs resulting in a greater number of categorized genes and less redundancy in gene classes. The most significant classes of genes were selected for further analyses, which can be seen in **Figure 5**. All the remaining classes in which the DEGs of all conditions tested were classified according to KOG annotation are presented in **Supplementary Figure S6**. In both cases, data is shown as the percentage of DEGs in relation to the total number of genes in the *R. toruloides* genome.

We used the transcriptomic data generated in this work to search for putative TFs and putative TFBS that might be playing a role in the regulation of both the stress response and the response to complex carbon sources. The pursuit of TF and TFBS was carried out in several stages in order to select suitable candidates. First, we extracted the sequence from the promoters corresponding to the genes of the KOG classes from **Figure 5**. By doing this, we both restricted the quantity of sequences to search motifs, which may reduce background noise, and we attempted to guarantee a biological relationship between these promoter sequences. Then, we used the HOMER software to obtain the putative motifs from those sequences. The motifs obtained were read on the

R platform using the universal motif library and a pairwise comparison was performed between all of them, with Pearson's correlation coefficient (PCC) as a method. The pairwise comparison was done following the rationale that if the motif is found in more than one gene, the chances of being a real motif increase. We also performed pairwise alignment using TOMTOM to complement the analysis. Finally, we filtered the results by a p-value less than 0.05, an e-value less than 1 and a PCC greater than 0.8. The resulting motifs were then compared to the Jaspar fungi 2018 (non-redundant DNA) database using TOMTOM, and some of these alignments can be seen in **Figure 6**. The motif candidates and motifs from JASPAR, aligned by TOMTOM, were plotted in a graph so that we could analyze the similarities between them (**Figure 7).** The main TFs corresponding to the putative TFBS found using TOMTOM are listed in **Table 2**. The percentage of identity of TF proteins found by this method against proteins from *R. toruloides* genome are also listed in **Table 2,** along with GO functions of the TF proteins.

4. Discussion

In this work, we developed a reproducible bioinformatics workflow that can be used for the discovery of regulatory elements using transcriptomic data as input, and applied to the emergent microbial cell factory *R. toruloides*. First, by analyzing the transcriptomic data, we showed that *R. toruloides* presented different transcriptional responses to all the conditions tested. When using media containing SCJ and urea, functional analysis showed a predominance of metabolism pathways being up-regulated, while degradation pathways such as autophagy and peroxisome are down-regulated. This pattern is reasonable since the yeast adapted its metabolism from the inoculum containing no sugar

to the rich and complex environment provided by SCJ. Interestingly enough, one of the pathways that were found to be up-regulated by GAGE was the terpenoid backbone biosynthesis. This is a good indicator that we can use the abundant and inexpensive media composed by sugarcane and urea as a start point substrate to produce terpenes in *R. toruloides,* like previously done using corn stover hydrolysates (Kirby *et al.*, 2021). Moreover, when growing the yeast at 42 ˚C, cytosolic DNA sensing pathways and mRNA surveillance pathways were found to be up-regulated and are representative of cell stress responses (Jamar, Kritsiligkou and Grant, 2017; Meng *et al.*, 2021). Similarly, in this condition, the homologous recombination pathway is being up-regulated – possibly for dealing with cell damaging (Litwin *et al.*, 2013). We also hypothesized that the transcriptional responses to the two high temperature conditions can be classified into two distinct types of stress responses. The 42 ˚C condition caused the heat-shock response, when cells repress protein biosynthesis – as we can see by the decrease of the biosynthesis of amino acids - and protein misfolding and oxidative stress occur (Verghese *et al.*, 2012). In contrast, the condition of 37 ˚C might have reached the other type of response: thermotolerance – given that most transcripts were not differentially expressed compared to the control grown at 30 ˚C. The very few pathways found to be enriched in this condition, like cell cycle and oxidative phosphorylation, might suggest that *R. toruloides* is enduring the increase in temperature (Shui *et al.*, 2015; Huang *et al.*, 2018). Nevertheless, another study showed that cultivating *R. toruloides* at 37 ˚C greatly impaired its growth (Wu *et al.*, 2020). Yet, the ethanol response greatly corresponds with what is described in literature for ethanol stress and tolerance in *S. cerevisiae* (Stanley *et al.*, 2010). It is interesting to note that cell cycle and DNA replication pathways are both

being down-regulated, showing inhibition of cell growth, as it was expected for ethanol stress (Stanley *et al.*, 2010). Inhibition of endocytosis was also reported as an ethanol stress response (Lucero *et al.*, 2000) and we can see that both concentrations of ethanol resulted in a decrease of expression of *R. toruloides* endocytosis genes. Additionally, the proteasome genes are up-regulated, demonstrating that cells are dealing with misfolded proteins, which is described in literature as a sign of ethanol stress (Bubis *et al.*, 2020). Hence, we hypothesized that the response to the two different concentrations of ethanol is still a stress response and *R. toruloides* has not yet been able to achieve tolerance to the compound (Lucero *et al.*, 2000; Stanley *et al.*, 2010; Bubis *et al.*, 2020).

Furthermore, we explored the cis-regulatory elements related to the DEGs found in the transcriptomic analysis. The TFBS predicted for *R. toruloides* presented similarity mainly with binding sites for the following TFs: ARG80, RPN4, ADR1 and DAL81, and those TFs can be seen in the graph **(Figure 7)** as having the greater number of similar putative motifs. As demonstrated in **Table 2**, HAC1 protein is involved in ER-unfolded protein response and RPN4 protein is involved in response to stress, which could explain the enrichment of both their binding motifs in our industrial stress conditions. In fact, studies showed that mutations in RPN4 protein conferred ethanol resistance to *S. cerevisiae* (Bubis *et al.*, 2020). However, there were no homologs of these proteins found in our carotenogenic yeast genome. Some TFs were found to be responsive to oxidative stress, such as AFT2 and HCM1, whereas only HCM1 was found in the *R. toruloides* genome. Interestingly, DAL81 protein was found to be mutated in response to high temperature stress in a study by Huang *et al.* (2018). This could explain the enrichment of its binding motifs in our heat shock transcripts, although a homolog of this protein was

not found. SWI6 binding motifs may be an interesting source for future studies, since this protein was found in *R. toruloides* genome with 39% identity and it is a protein that regulates heat stress response. Interestingly, binding motifs for SPT23, a protein that regulates cold response, were also found. This protein was found in the genome of *R. toruloides* with 45% identity. Nevertheless, the binding of the TF proteins to their respective motifs discovered by our pipeline needs to be further tested to confirm their activity. Alternatively, in the case of the ones whose corresponding binding protein homologs were not found, recognition of the identified motifs could be performed by yet unknown TFs. As shown by Antonieto *et al.*, (2019), a transcriptional regulator with low homology to AZF1 recognizes a well-conserved motif regulating expression of cellulases genes.

The motifs found can be used as starting points for new transcriptional modulation studies. The ultimate goal would be to define a core promoter sequence for *R. toruloides*, and then change the sequences of TFBS according to the desired behavior. For instance, one can study ethanol tolerance in this organism by adding the RPN4 consensus sequence to their respective constructs. Additionally, our pipeline can be applied to other organisms that already have well-defined core promoter sequences, where scientists can "play around" with both motif sequences and locations in order to unlock new transcriptional behaviors (Monteiro *et al.*, 2020; Tominaga, Kondo and Ishii, 2022). Also, as mentioned in the Results section, our transcriptomic data can be used to create new promoter libraries for our yeast. Ultimately, anyone can resort to our tools to generate better transcriptional understanding of their organism of interest.

5. Conclusion

Here, we characterized the transcriptional responses of *R. toruloides* when growing in industry-like conditions, and discovered new regulatory elements that were enriched in each context. We showed how differential gene expression, followed by a custom, reproducible bioinformatics motif discovery workflow, was able to predict putative motifs for binding of transcription factors in our strain, most of them involved in stress-related responses. Using our approach, DNA motifs similar to the binding sites for ADR1, ARG80, DAL81 and RPN4 transcription factors were found to be the most abundantly enriched in our dataset. While ADR1 and ARG80 proteins were found in *R. toruloides* genome with 54% and 60% of identity, respectively, DAL81 and RPN4 were not found. The novel putative *cis*-regulatory elements described here offer a great initial point to optimize gene regulation in industrial conditions expanding the current knowledge on regulatory networks for this yeast. Furthermore, our pipeline for motif discovery can be easily applied for other unexplored hosts.

6. Figures

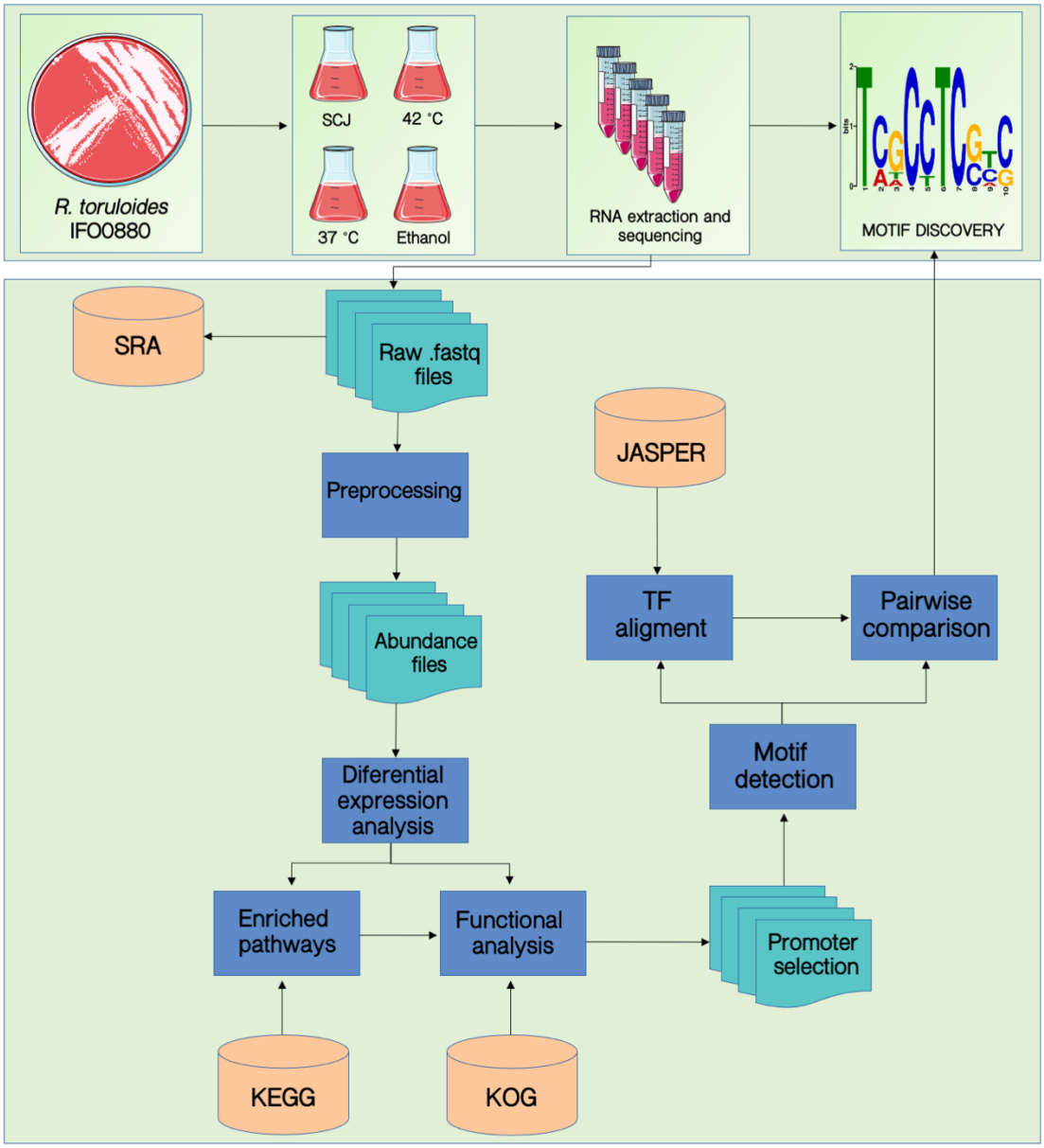


Figure 1. Workflow for the discovery of *cis*-regulatory elements using transcriptomic data.

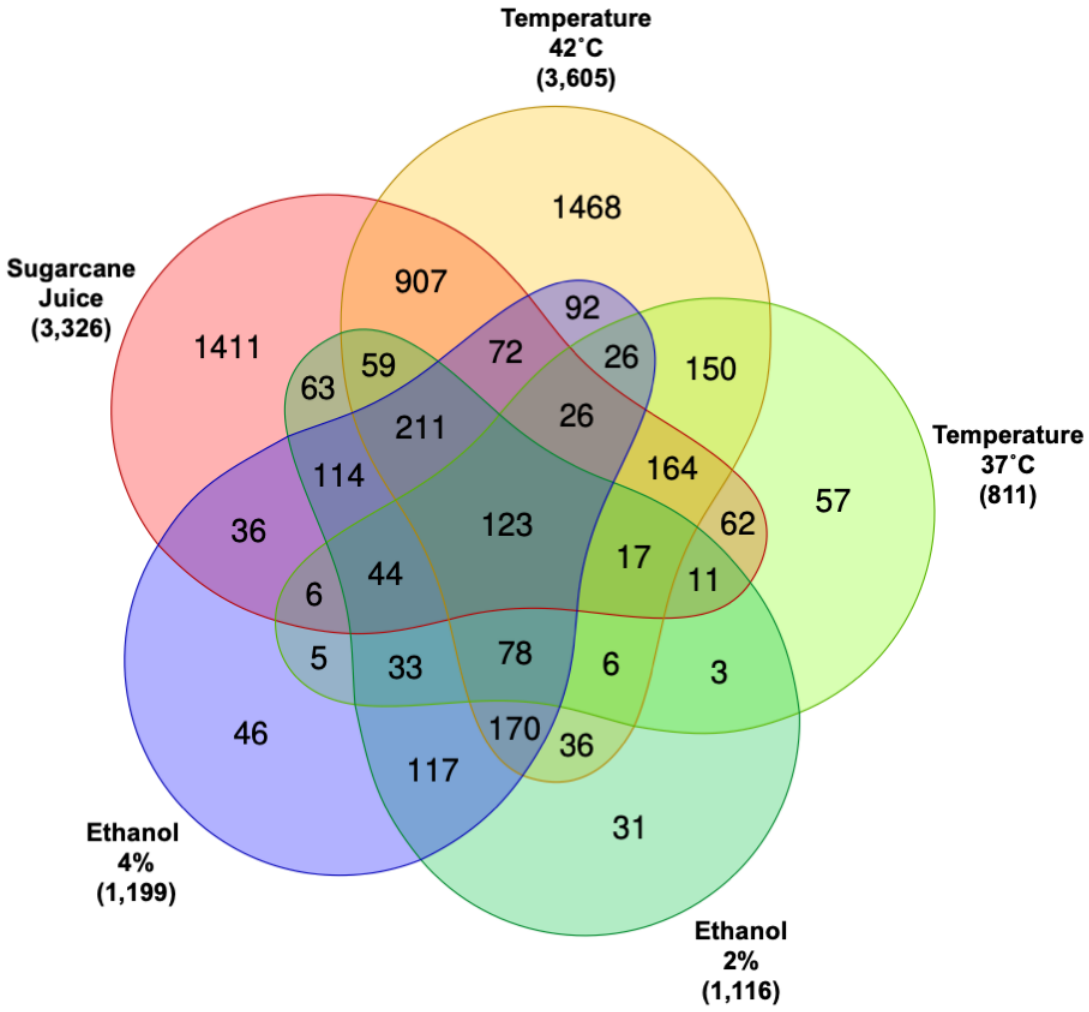


Figure 2. Venn Diagram of DEGs for all conditions. The Venn Diagram shows how many DEGs are shared between all conditions and how many are exclusive for each condition. Numbers outside of the Venn diagram represent the total of DEGs when compared to each respective control.

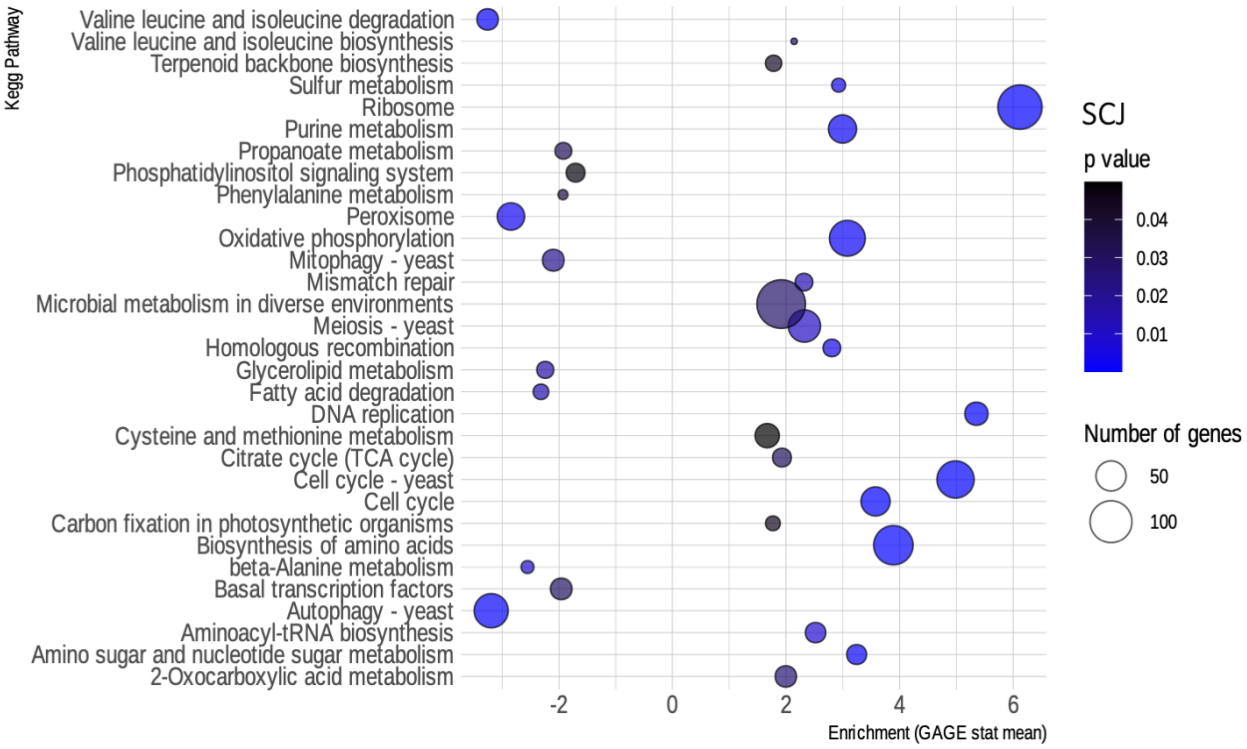


Figure 3. Enriched KEGG pathways for *R. toruloides* grown on SCJ. Bubble map showing the biochemical pathways of *R. toruloides* noted by KEGG enriched in the SCJ condition, as obtained by the GAGE package. Pathways that have an enrichment value greater than 0 are up-regulated, while those that have a value less than 0 are down-regulated. Blue scale inside the bubbles represents the decreasing *p*-values. The different sizes of the bubbles define the approximate number of DEGs in each biochemical pathway.

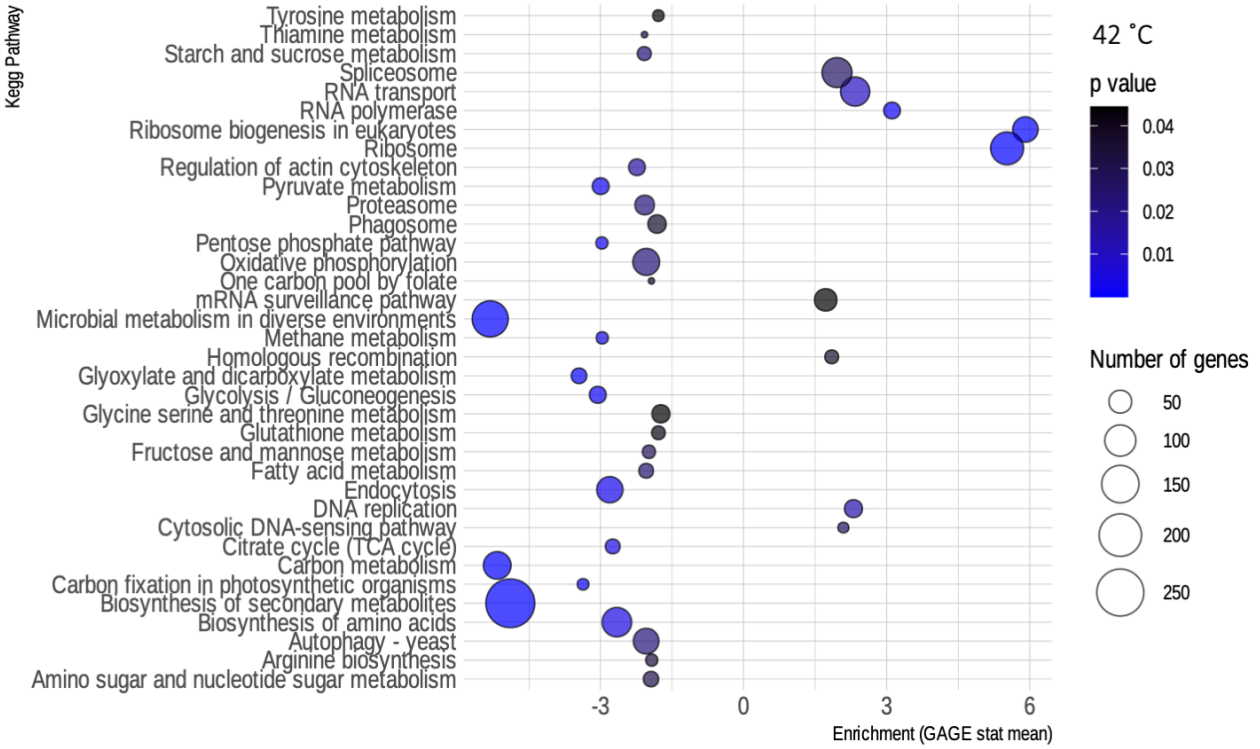


Figure 4. Enriched KEGG pathways for *R. toruloides* grown at 42 ˚C. Bubble map showing the biochemical pathways of *R. toruloides* noted by KEGG that are enriched in the 42 ˚C condition, as obtained by the GAGE package. Pathways that have an enrichment value greater than 0 are up-regulated, while those that have a value less than 0 are down-regulated. Blue scale inside the bubbles represents the decreasing p-values. The different sizes of the bubbles define the approximate number of DEGs in each biochemical pathway.

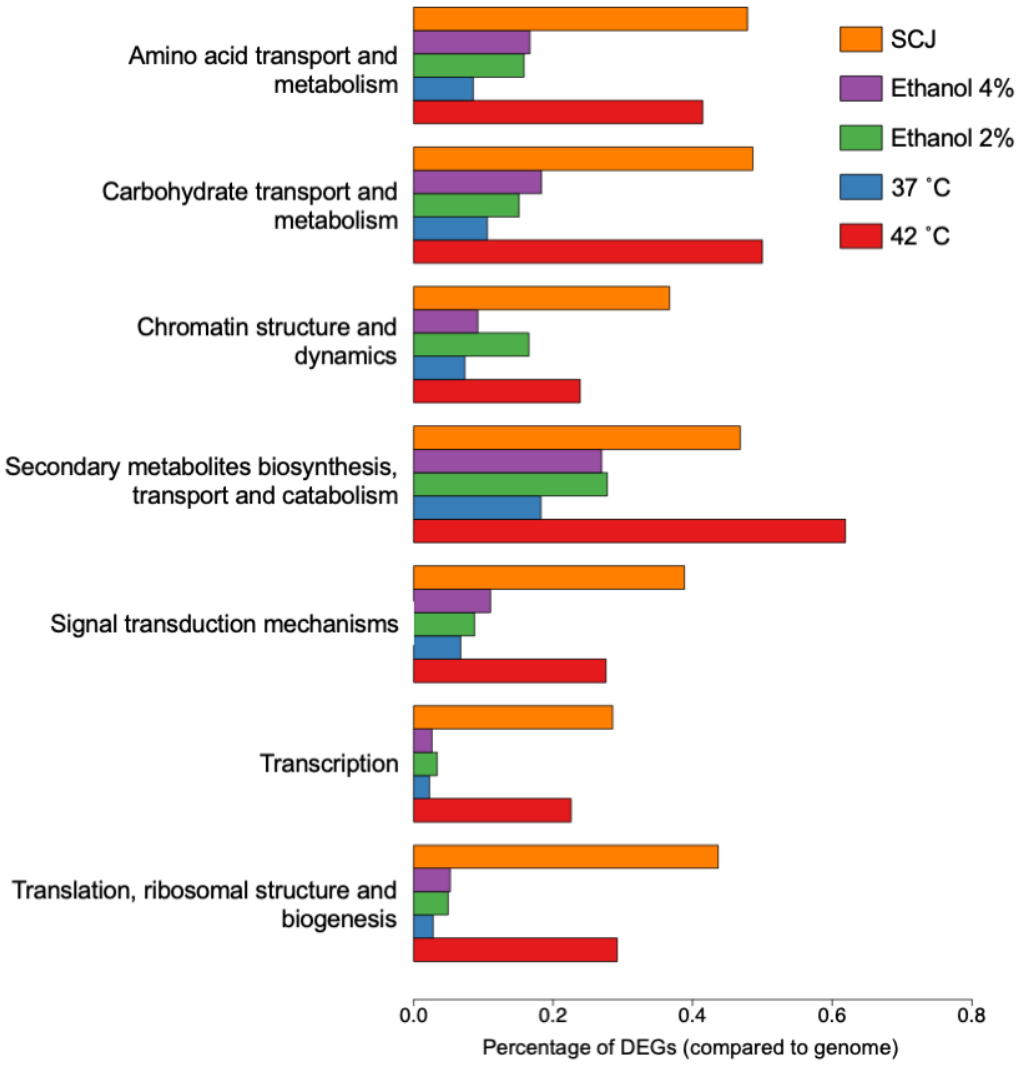


Figure 5. KOG groups selected for TFBS finding. Percentage of DEGs compared to the total number of genes in the *R. toruloides* genome for each condition are shown as annotated using KOG. These are the most relevant conditions with the greatest amount of DEGs and were selected for further analysis. The remaining KOG annotations are shown in **Supplementary Figure 5**.

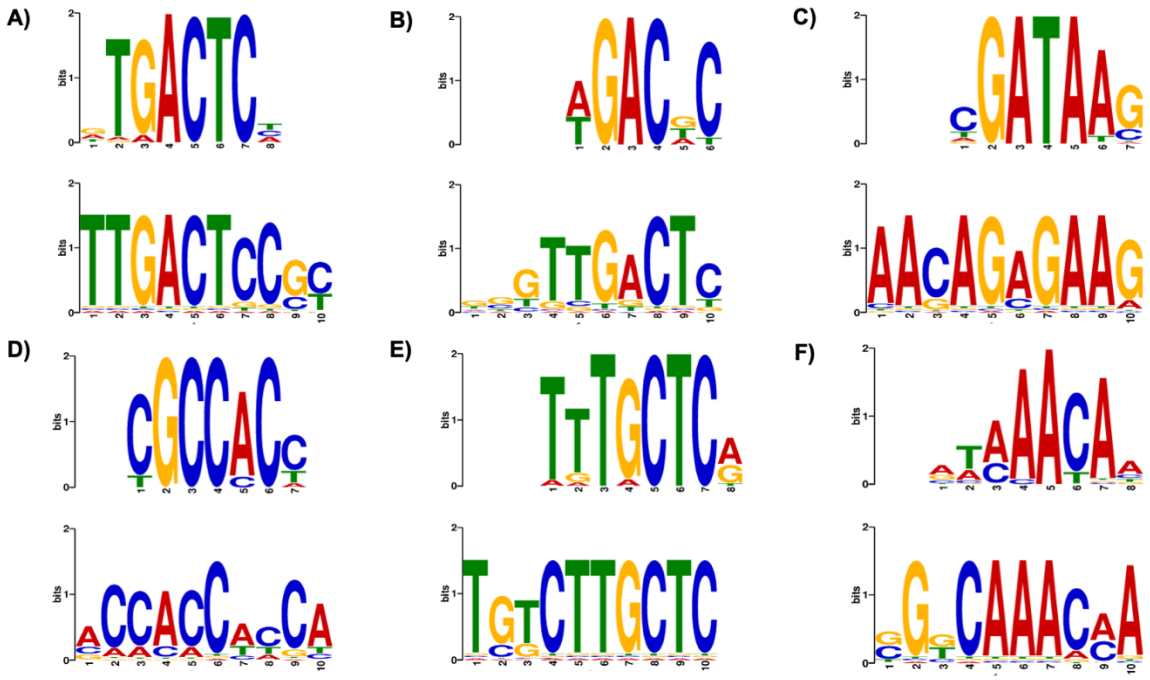


Figure 6. Examples of putative motifs aligned with transcription factors motifs from the JASPAR database using TOMTOM. (A) Top motif is the binding motif for ARG81 from *S. cerevisiae*. Bottom motif is from SCJ condition. KOG class: Amino acid transport and metabolism. *p*-value: 1.08e-03. *E*-value: 1.51e+00 **(B)** Top motif for AGR80 from *S. cerevisiae*. Bottom motif is from 42 ˚C condition. KOG class: Amino acid transport and metabolism. *p*-value: 4.98e-03. *E*-value: 8.76e-01. **(C)** Top motif for DAL80 from *S. cerevisiae*. Bottom motif is from ethanol 2% condition. KOG class: Amino acid transport and metabolism. *p*-value: 3.38e-02. *E*-value: 5.95e+00. **(D)** Top for RPN4 from *S. cerevisiae*. Bottom motif is from 42 ˚C condition. KOG class: Chromatin structure and dynamics. *p*-value: 4.15e-03. *E*-value: 7.31e-01. **(E)** Top motif for MAC1 from *S. cerevisiae*. Bottom motif is from ethanol 2% condition. KOG class: Secondary metabolites biosynthesis transport and catabolism. *p*-value: 4.15e-03. *E*-value: 7.30e-01. **(F)** Top motif for HCM1 from *S. cerevisiae*. Bottom motif is from SCJ condition. KOG class: Chromatin structure and dynamics. *p*-value: 3.44e-02. *E*-value: 6.05e+00.

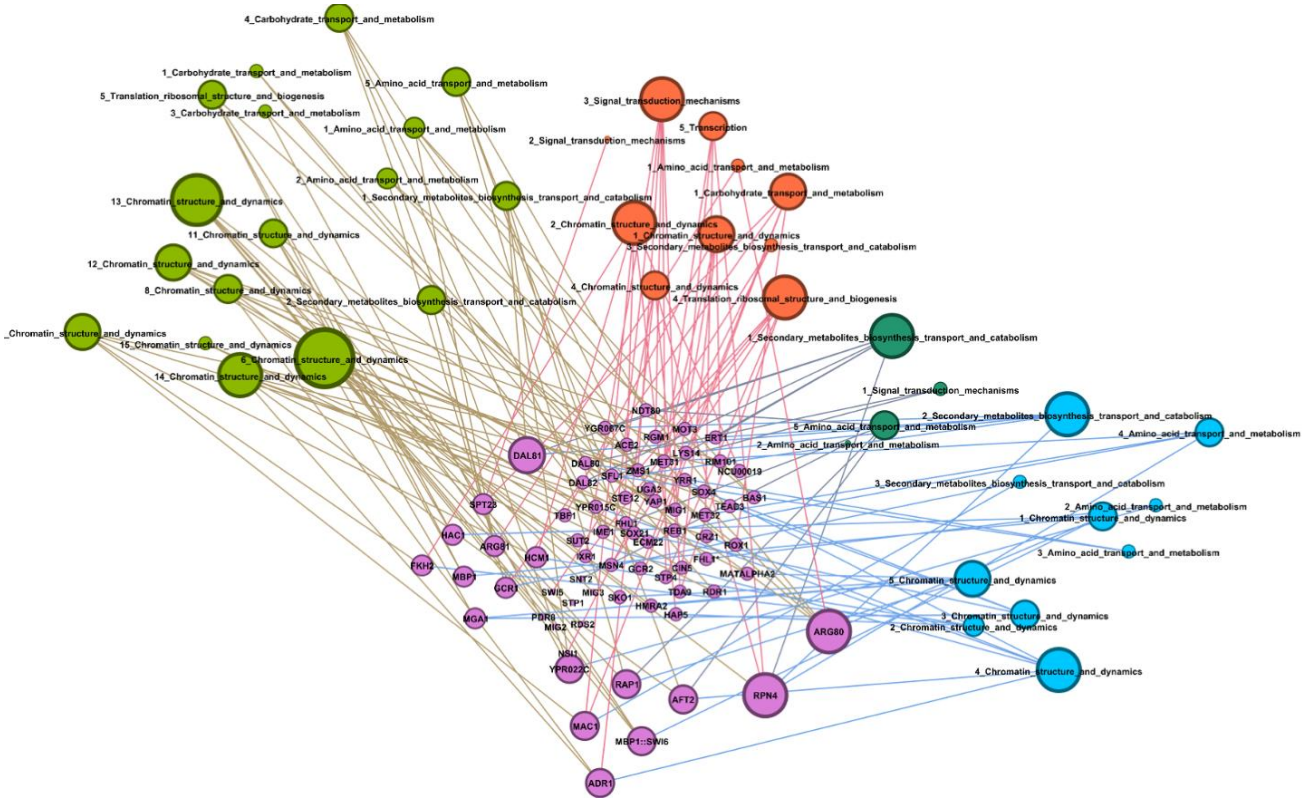


Figure 7. Graph of putative transcription factors and their putative binding sites. Graph showing interaction between putative transcription factors for *R. toruloides* with the TF binding motifs predicted using HOMER. Green nodes represent the sugarcane juice condition, orange nodes represent the 42 ˚C condition, dark green nodes are the 2% ethanol, and blue nodes are the 4% ethanol condition. Written inside nodes are the number of the TFBS and the name of the gene class annotated by KOG in which that motif was found. Pink nodes are the putative transcription FTs found using TOMTOM.

7. Tables

Table 1. **Mining of promoter and TFBS sequences.** Number of promoter sequences selected for each condition and each KOG class and the number of possible motifs found by HOMER.

Condition	KOG Class	Promoters	Motifs
SCJ	Amino acid transport and metabolism	111	5
	Carbohydrate transport and metabolism	105	5
	Chromatin Structure and dynamics	39	16
	Secondary metabolites biosynthesis, transport and catabolism	59	4
	Signal Transduction Mechanisms	134	4
	Transcription	76	5
	Translation, ribosomal structure and biogenesis	140	5
Temperature 37ºC	Amino acid transport and metabolism	20	2
	Carbohydrate transport and metabolism	23	3
	Chromatin Structure and dynamics	8	5
	Secondary metabolites biosynthesis, transport and catabolism	23	3
	Signal Transduction Mechanisms	23	2
	Transcription	6	5
	Translation, ribosomal structure and biogenesis	8	5
Temperature 42ºC	Amino acid transport and metabolism	97	4
	Carbohydrate transport and metabolism	106	5
	Chromatin Structure and dynamics	26	5
	Secondary metabolites biosynthesis, transport and catabolism	76	5
	Signal Transduction Mechanisms	96	3
	Transcription	60	5
	Translation, ribosomal structure and biogenesis	93	5
	Amino acid transport and metabolism	37	4

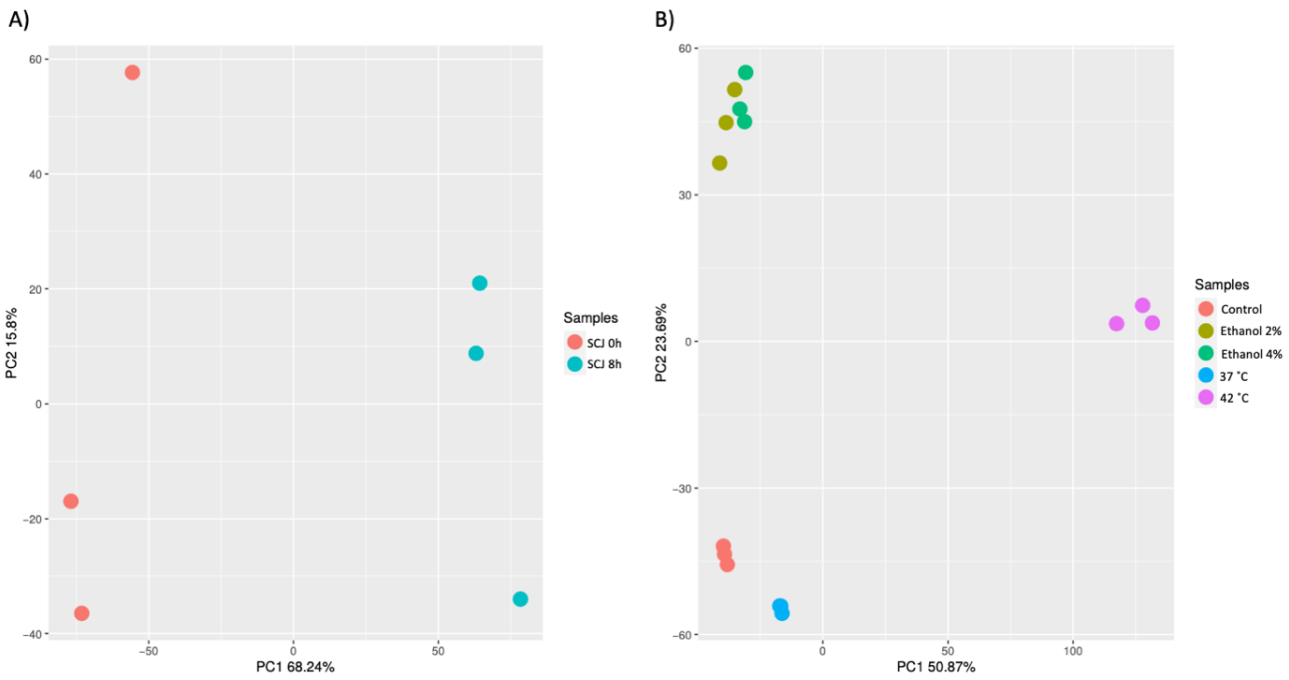
Ethanol 2%	Carbohydrate transport and metabolism	32	0
	Chromatin Structure and dynamics	18	5
	Secondary metabolites biosynthesis, transport and catabolism	35	5
	Signal Transduction Mechanisms	30	2
	Transcription	9	0
	Translation, ribosomal structure and biogenesis	15	0
Ethanol 4%	Amino acid transport and metabolism	39	5
	Carbohydrate transport and metabolism	39	5
	Chromatin Structure and dynamics	10	0
	Secondary metabolites biosynthesis, transport and catabolism	34	4
	Signal Transduction Mechanisms	39	4
	Transcription	7	0
	Translation, ribosomal structure and biogenesis	17	0

Table 2. **Predicted transcription factors and their sequences.** BLAST of the predicted transcription factors found by comparison of putative motifs against *R. toruloides* genome showing the percentage of identity for each protein sequence. Protein sequences and Gene Ontology numbers extracted from the Uniprot database. Only some predictive GO functions are shown, according to its correspondence to the study scope.

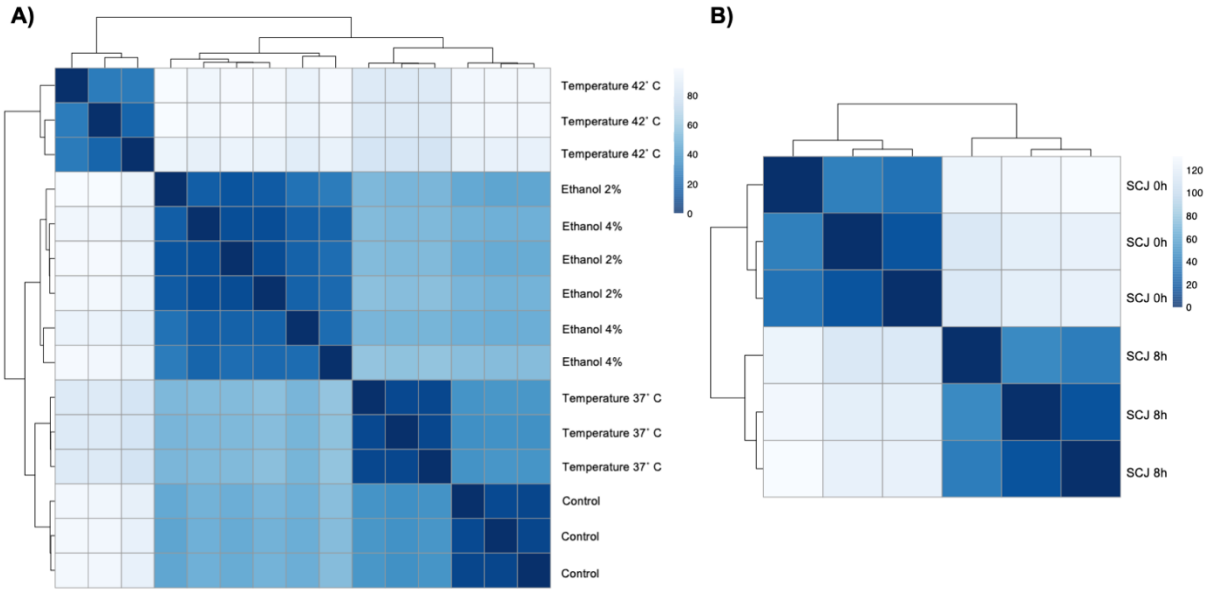
TF	GO Number	GO Function	Found in *R. toruloides*?	% of identity
ADR1	GO:0061410	Positive regulation of transcription in response to ethanol	Yes	54%
AFT2	GO:0034599	Any process that results in a change in state of a cell as a result of oxidative stress	No	-
ARG80	GO:0006525	Chemical reactions and pathways involving arginine	Yes	60%
ARG81	GO:0006525	Chemical reactions and pathways involving arginine	No	-
DAL81	GO:1901717	Activates or increases the frequency, rate or extent of gamma-aminobutyric acid catabolic process	No	-
FKH2	GO:0006338	Dynamic structural changes to eukaryotic chromatin throughout the cell division cycle	Yes	47%
GCR1	GO:0000433	Carbon catabolite repression of transcription from RNA polymerase II promoter by glucose	No	-
HAC1	GO:0030968	Endoplasmic reticulum unfolded protein response	No	-
HCM1	GO:0034599	Any process that results in a change in state of a cell as a result of oxidative stress	Yes	49%
MAC1	GO:0045732	Positive regulation of protein catabolic process	Yes	50%

MBP1	GO:0071931	Positive regulation of transcription involved in G1/S transition of mitotic cell cycle	Yes	52%
MGA1	GO:0007124	Pseudohyphal growth in response to an environmental stimulus	Yes	51%
RAP1	GO:0010833	Telomere maintenance via telomere lengthening	No	-
RPN4	GO:0097201	Negative regulation of transcription from RNA polymerase II promoter in response to stress	No	-
SPT23	GO:0070417	Cellular response to cold	Yes	45%
SWI6	GO:0061408	Positive regulation of transcription from RNA polymerase II promoter in response to heat stress	Yes	39%
YPR022C	GO:0000981	Modulates the transcription of specific gene sets transcribed by RNA polymerase II	Yes	69%

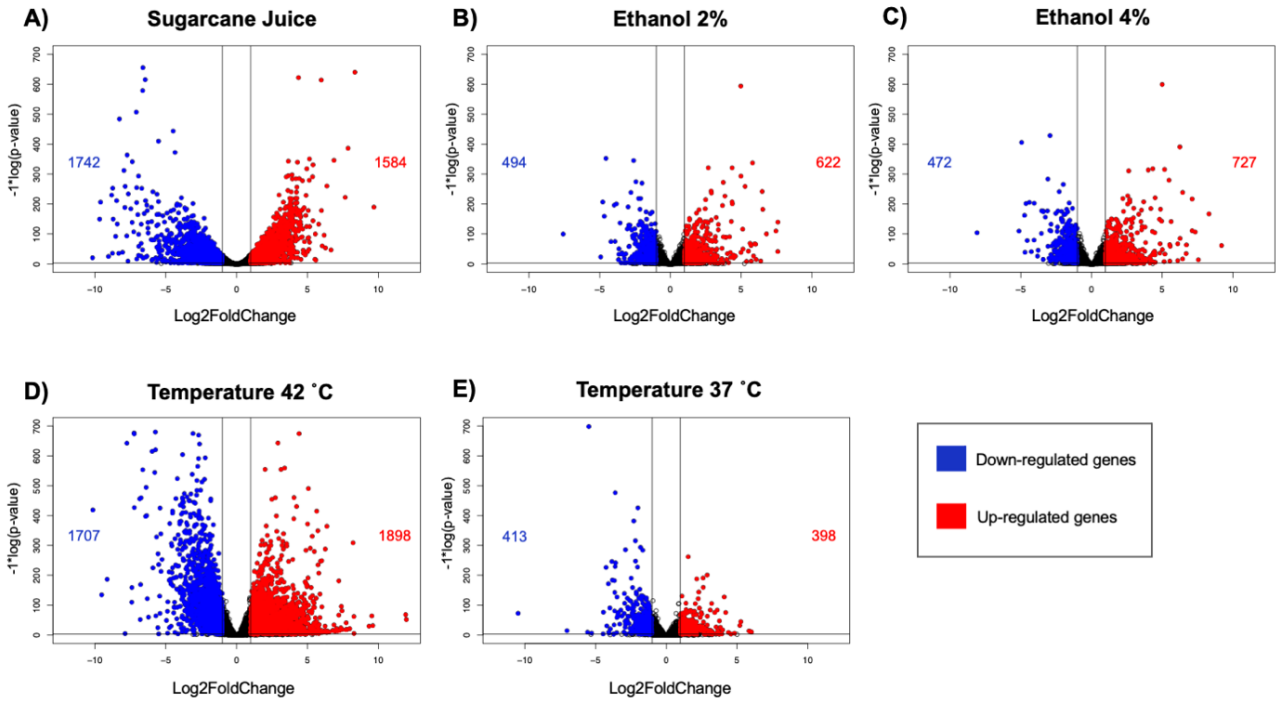
8. Supplementary Material



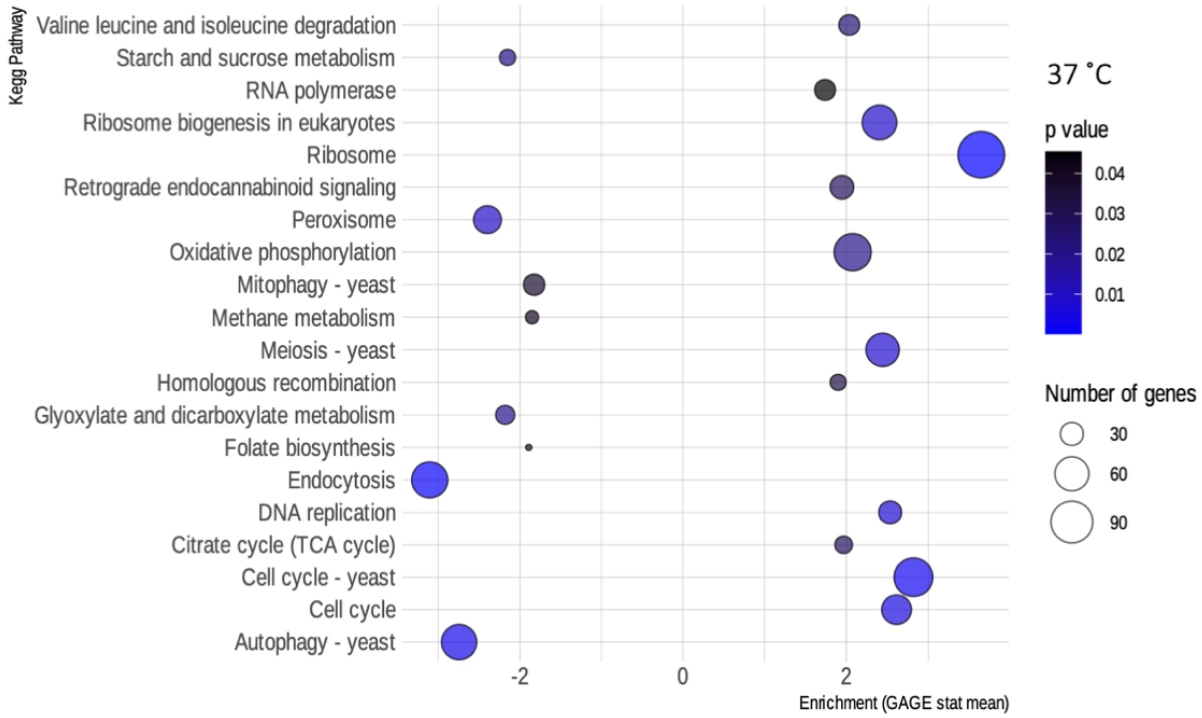
Supplementary Figure S1. PCA of experimental replicates. Principal Component Analysis (PCA) applied to *R. toruloides* transcripts, grouping the conditions based on the similarity of the projection. **(A)** Experiment using SCJ as a substrate. The conditions of this experiment are time 0 h (control cultures grown in LB for 24 hours) and 8 h (cultures transferred from the pre-inoculum to a medium containing sugarcane juice and urea). **(B)** Experiment using industrial stress conditions. The conditions for this experiment are control time (grown in YPD for 24 hours) and the respective stress conditions: 2% ethanol, 4% ethanol, 42 ˚C and 37 ˚C.



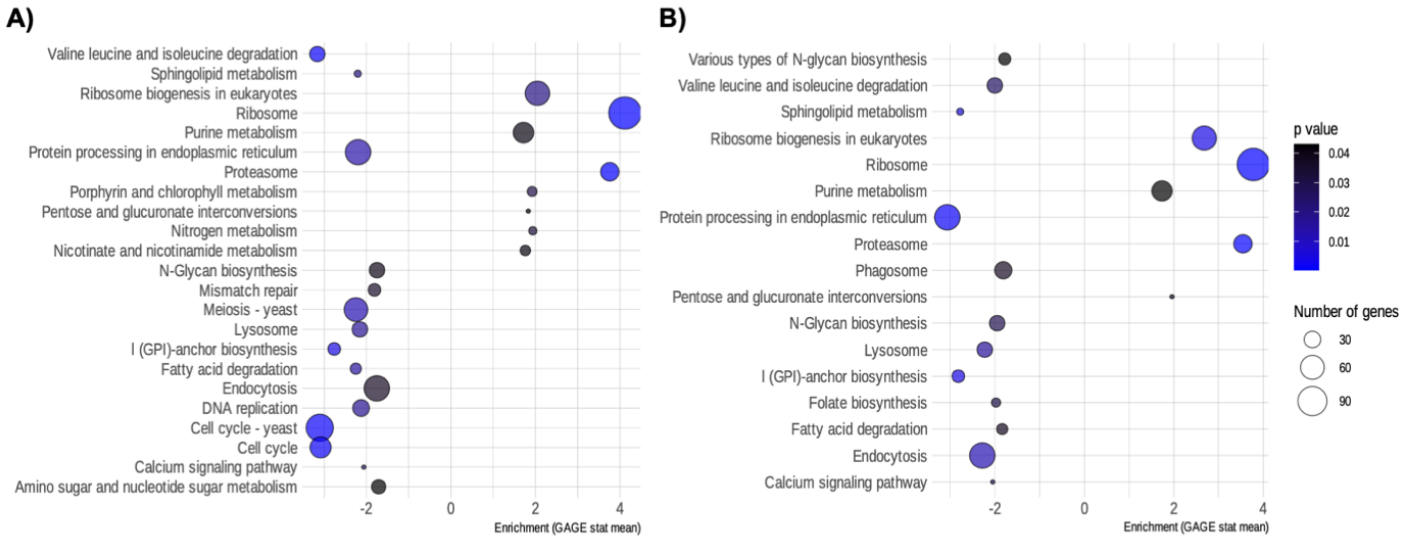
Supplementary Figure S2. Heatmap for experimental replicates measured by Euclidean Distance. (A) Heatmap showing the correlation of gene expression between transcripts of *R. toruloides* grown under conditions of industrial stress. The conditions for this experiment are control (grown in YPD for 24 hours) and the respective stress conditions: ethanol 2%, ethanol 4%, 42 ˚C and 37 ˚C. **(B)** Heatmap showing the correlation of gene expression between transcripts of *R. toruloides* grown in SCJ. The scale represents Euclidean distance, where the lighter the blue, the greater the distance between the samples.



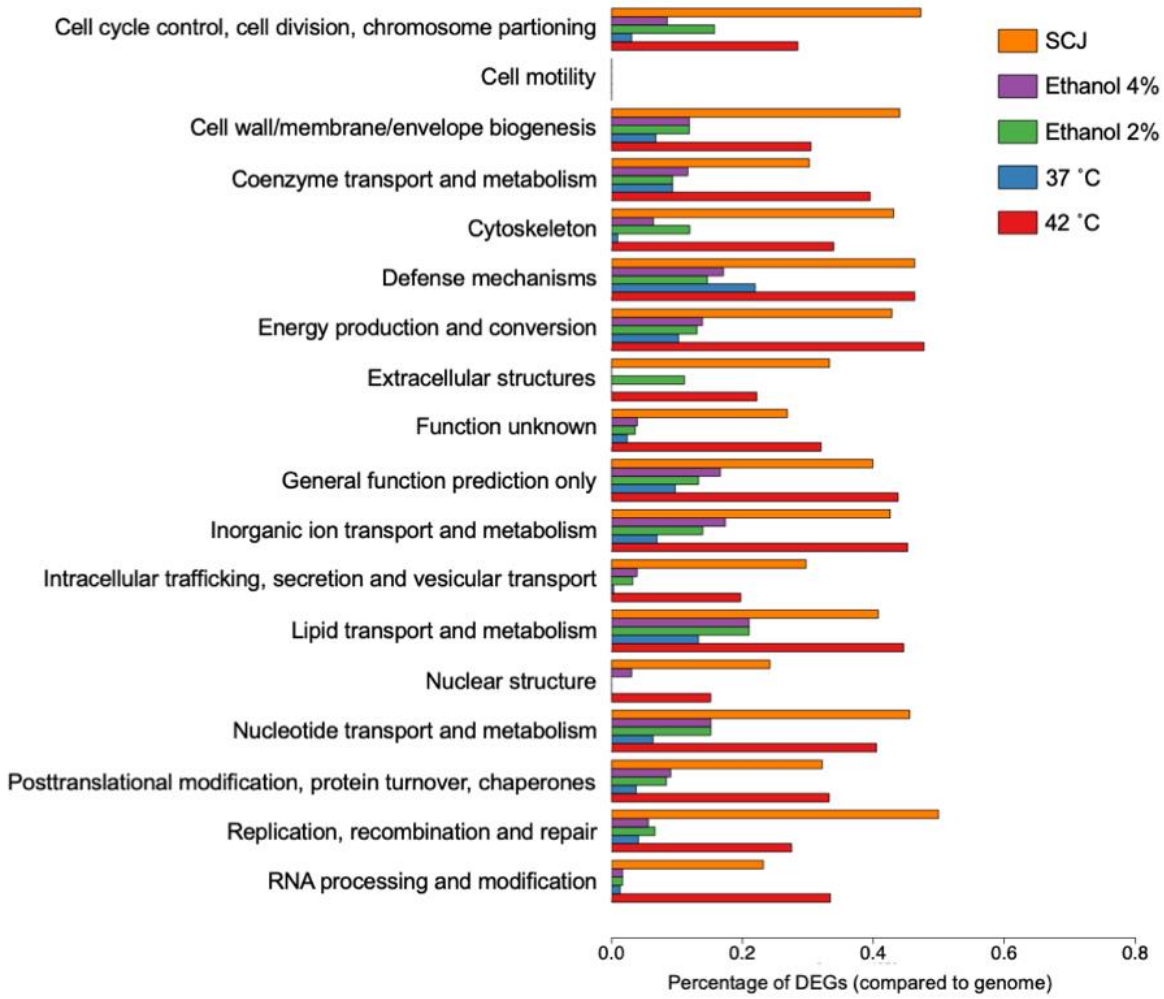
Supplementary Figure S3. Volcano plots for *R. toruloides* DEGs in all conditions. Volcano plot showing the DEGs for each experimental setting when compared to its respective control. Genes that are down-regulated are represented in blue and genes that are up-regulated are represented in red. The vertical cut lines on the graph divide the log2FoldChange values less than -1 (left) and greater than 1 (right). The horizontal cut line in the graph divides the valid p-value values less than 0.05 in -log (*p*-value).



Supplementary Figure S4. Enriched KEGG pathways for *R. toruloides* grown at 37 ˚C. Bubble map showing the biochemical pathways of *R. toruloides* noted by KEGG that are enriched in the 37 ˚C condition, as obtained by the GAGE package. Pathways that have an enrichment value greater than 0 are up-regulated while those that have a value less than 0 are down-regulated. Blue scale inside the bubbles represents the decreasing *p*-values. The different sizes of the bubbles define the approximate number of DEGs in each biochemical pathway.



Supplementary Figure S5. Enriched KEGG pathways for *R. toruloides* grown in ethanol conditions. Bubble map showing the biochemical pathways of *R. toruloides* noted by KEGG that are enriched in the ethanol conditions, as obtained by the GAGE package. **(A)** Ethanol 2%. **(B)** Ethanol 4%. Pathways that have an enrichment value greater than 0 are up-regulated while those that have a value less than 0 are down-regulated. Blue scale inside the bubbles represents the decreasing *p*-values. The different sizes of the bubbles define the approximate number of DEGs in each biochemical pathway.



Supplementary Figure S6. Percentage of DEGs annotated using KOG compared to the total number of genes in the *R. toruloides* genome. Percentage of DEGs compared to the total number of genes in the *R. toruloides* genome for each condition are shown as annotated using KOG.

CHAPTER TWO

Metabolic engineering for monoterpene production in *Rhodosporidium toruloides*

ABSTRACT

Robust microbial hosts that produce value-added chemicals are of extreme importance, and *Rhodosporidium toruloides* is considered one of the most promising oleaginous yeasts for industrial applications. The aim of this work consisted in the application of metabolic engineering methods to create *R. toruloides* strains that are capable of producing non-native terpenes. Terpenes are important molecules with several applications such as food additives, fragrances, and biofuels. Three monoterpene synthase enzymes were chosen for heterologous expression of monoterpenes of commercial interest. The pinene synthase from *Abies grandis* was expressed to produce pinene, which can compose high-density renewable aviation fuels. Linalool synthase from *Mentha citrate* and geraniol synthase from *Ocimum basilicum* were expressed to produce beer flavorings without hops. A truncated form of the HMG1 enzyme was also expressed to increase the flux through the mevalonate pathway. Constructs expressing these genes were successfully assembled using Golden Gate Assembly and transformed into *R. toruloides*. Positive transformants were obtained for all genes, confirmed by either PCR or qPCR. When no monoterpenes were detected in the first mutant strains, new rounds of Design, Build, Test were conducted by inserting more copies of the expression cassettes. This work followed the rationale of the DBTL cycle developed for synthetic biology, where constructs were designed to produce non-native terpenes, built using a modular Golden Gate assembly strategy, and tested in different conditions, including flasks and bioreactors.

Keywords: pinene, linalool, geraniol, mevalonate pathway, heterologous expression.

Abbreviations:

BLE - Bleomycin

CAM - Chloramphenicol

DBTL - Design, build, test, learn

FPP - Farnesyl diphosphate

FPPS - Farnesyl diphosphate synthase

GES - Geraniol synthase

GGPP - Geranylgeranyl diphosphate

GPP - Geranyl diphosphate

GPPS - Geranyl diphosphate synthase

HMG1 - HMG-CoA reductase

HYG - Hygromycin

LIS - Linalool synthase

MS - Monoterpene synthase

MEP - Methylerythritol-phosphate pathway

MVA - Mevalonate

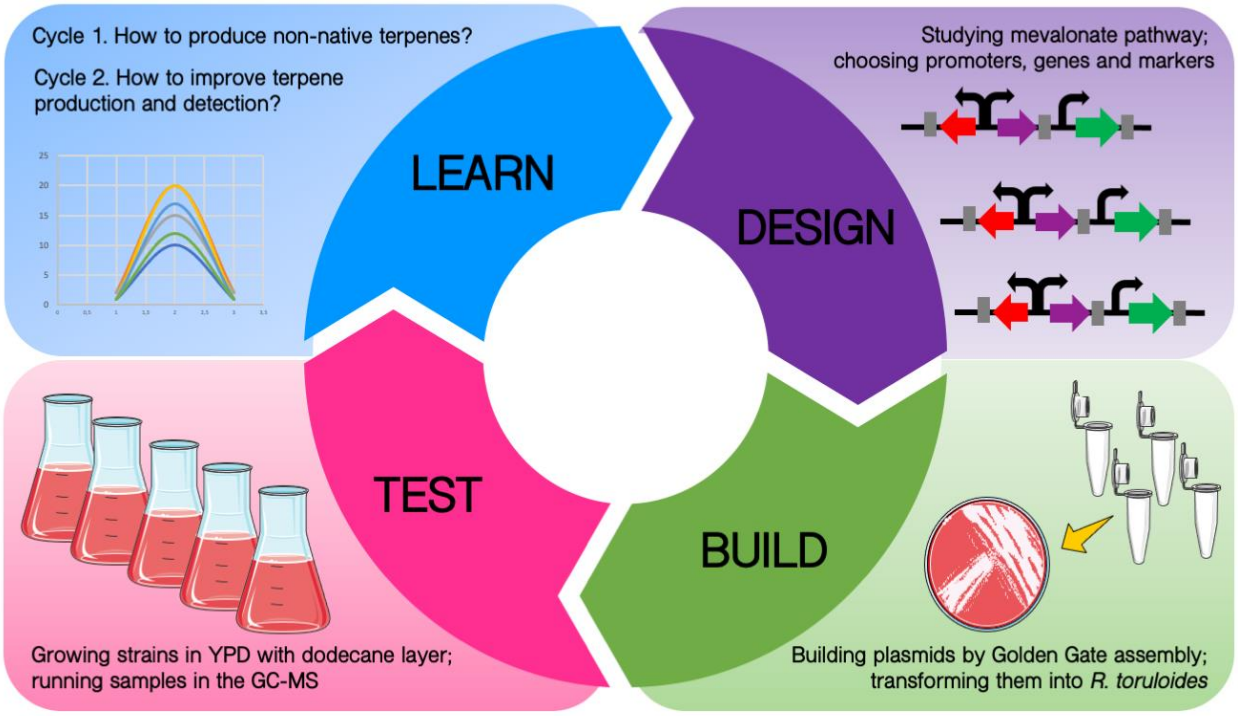
NAT - Nourseotrhicin

NDPS1 - Neryl diphosphate synthase

NPP - Neryl diphosphate

PS - Pinene synthase

GRAPHICAL ABSTRACT



1. Background

In the pursuit of a circular bio-based economy to obtain products without generating waste, the search for microorganisms that can turn residues into valuable chemicals is growing (Tan and Lamers, 2021; Y. Liu *et al.*, 2021). Microbial conversion of substrates has several advantages: productivity can be increased by optimization of microbes, as well as competition with food can be reduced, since biomass from agroindustrial residues can be used as starting material (Abomohra *et al.*, 2016; Bonturi *et al.*, 2017). *Rhodosporidium toruloides* is an emergent host that has already demonstrated an excellent capacity for application in microbial conversion. One of the main reasons is the broad range of substrates this yeast can utilize, including several different types of lignocellulosic biomass, like sugarcane bagasse and wood hydrolysates (Park, Nicaud and Ledesma-Amaro, 2017; Pinheiro *et al.*, 2020; Monteiro de Oliveira *et al.*, 2021). Furthermore, this yeast can accumulate high concentrations of lipids and has the potential to store other compounds derived from acetyl-CoA. Several studies have applied metabolic engineering in *R. toruloides* to increase the carbon flux towards acetyl-CoA-derived products to obtain fine chemicals, such as terpenes (Yaegashi, Kirby, Ito, Sun, Dutta, Mirsiaghi, Eric R. Sundstrom, *et al.*, 2017; Geiselman *et al.*, 2020; Kirby *et al.*, 2021).

Terpenes are composed of various isoprenoid chains and are widely used today for various purposes, including colorants, coatings, pharmaceuticals, fragrances, additives for food and cosmetics, and biofuels (Monteiro and Veloso, 2004; Paduch *et al.*, 2007; Kirby *et al.*, 2021). These compounds can be produced by two biosynthetic pathways, the cytosolic mevalonate (MVA) pathway and the methylerythritol-phosphate

(MEP) pathway, whereas *R. toruloides* utilizes the MVA pathway to produce its native isoprenoids (Wen *et al.*, 2020; Li *et al.*, 2022). Therefore, the remarkable capacity of *R. toruloides* to accumulate acetyl-CoA attracts the use of metabolic engineering strategies targeting the pathways producing these compounds (Park, Nicaud and Ledesma-Amaro, 2017; Kirby *et al.*, 2021; Li *et al.*, 2022). Based on the rationale that the production of fine chemicals using microbial conversion will have a low cost, a more reliable and abundant supply and a more ecologically sustainable approach (Martins-Santana *et al.*, 2018), the aim of this work was to combine synthetic biology and metabolic engineering strategies to engineer the oleaginous yeast *R. toruloides* to produce terpenes, more specifically, monoterpenes.

Monoterpenes are terpenes with ten carbons that use GPP as precursors. They are volatile and are commonly found in plants, serving critical roles in the interactions of plants with their environments (Jiang, Kempinski and Chappell, 2016). Some studies have already demonstrated the production of non-native monoterpenes in other microorganisms, such as *Escherichia coli, Saccharomyces cerevisiae* and *Yarrowia lipolytica* (Sarria *et al.*, 2014; Niu *et al.*, 2018; Wei *et al.*, 2021; Li *et al.*, 2022), and also for *R. toruloides* (Zhuang *et al.*, 2019; Kirby *et al.*, 2021; S. Liu *et al.*, 2021). Three important monoterpenes were chosen to be produced in this study: pinene, linalool and geraniol. Pinenes can be used for many different applications, from fragrances to anti-inflammatory drugs, and mainly as part of the composition of high-density renewable aviation fuels (Sarria *et al.*, 2014; Wei *et al.*, 2021; Woodroffe *et al.*, 2021). Linalool and geraniol have applications as fragrances. In this specific case, the most significant interest is flavoring beers without using hops, as demonstrated previously (Denby et al., 2018).

The conversion from the substrate GPP to monoterpenes requires only one reaction carried on by enzymes called monoterpene synthases (MS). For this reason, we chose our three suitable monoterpene synthases from plants. The pinene synthase (PS) sequence from the tree *Abies grandis*, based on the work by Sarria et al. (2014), where they tested several pinene synthases in *E. coli* and found this one to perform better. Several studies have successfully produced geraniol and linalool using *S. cerevisiae* as a host (Oswald *et al.*, 2007; Pardo *et al.*, 2015; Denby *et al.*, 2018; Zhou *et al.*, 2020; R. Li *et al.*, 2021). Based on those studies, the geraniol synthase (GES) sequence was extracted from basil (*Ocimum basilicum*) and linalool synthase (LIS) from mint (*Mentha citrata*).

Thus, constructs for expression of these genes were designed and successfully assembled using a Golden Gate Assembly method specially developed for application in *R. toruloides* (Bonturi, Pinheiro, de Oliveira, *et al.*, 2022), modulated by promoters characterized by Nora et al (2019), including a bidirectional one. After the first round of the DBTL cycle, we developed new constructs for overexpression of the genes of interest, aiming at integrating more copies into the genome. They were transformed into *R. toruloides* and positive transformants were confirmed by PCR. Mutant strains were obtained for all three targets, ready for characterization. Detection of monoterpene production in the new mutant strains remains to be shown, confirming that we successfully applied synthetic biology and metabolic engineering strategies to produce monoterpenes in the emergent, oleaginous yeast *R. toruloides*.

2. Materials and Methods

2.1. Strains and growth conditions

The parental *R. toruloides* strains used in this work were CCT7815 and IFO0880. *R. toruloides* IFO0880 was used because of its haploidy (Coradetti *et al.*, 2018). The strain CCT7815 is named SBY29 on the laboratory registry and will be called SBY29 from now on. This strain is derived from *R. toruloides* CCT0783 after a short-term adaptation in sugarcane bagasse and is a more robust strain (Bonturi *et al.*, 2017). Unless specified, *R. toruloides* was grown in either liquid or solid YPD medium (glucose, 20 g/L; yeast extract, 10 g/L; peptone, 20 g/L). The *E. coli* strain used for all cloning procedures and plasmid propagation was DH5α maintained in either liquid or solid LB medium (tryptone, 10 g; NaCl, 10 g; yeast extract, 5 g). For solid medium, 20 g/L of agar was added. When necessary, antibiotics were added to the media depending on the marker of the respective construct, which includes: hygromycin (HYG) at 50 mg/mL, bleomycin (BLE) at 100 mg/mL, G418 at 200 mg/mL and nourseothricin (NAT) at 100mg/mL. For *E. coli* strains containing pGGA, the antibiotic added to the media was chloramphenicol (CAM) at a concentration of 25 mg/L.

2.2. Plasmids, parts and Golden Gate Assembly

All heterologous genes had their codons optimized for expression in *R. toruloides* and were synthesized by Twist Bioscience (USA). Except for the synthesized genes, all other constructs' components, such as promoters and terminators, were amplified by PCR from the yeast genome or from the plasmid collection. Sequences for constitutive

promoters used in all constructs were based on the library developed by Nora et al., (2019) (strains and plasmids also available at Joint BioEnergy Institute Strain Registry (https://public-registry.jbei.org/). All plasmids were built using the Golden Gate assembly platform dedicated to *R. toruloides* (Bonturi et al., 2022) using NEBridge® Golden Gate Assembly Kit (New England Biolabs) and transformed into chemically competent *E. coli* cells by heat-shock (Sambrook, J.; Fritsch, E. F.; Maniatis, 1989). These pGGA plasmids were recovered by miniprep and confirmed by sequencing in the Institute of Genomics Core Facility University of Tartu, in Tartu, Estonia. Once the expression cassettes were confirmed, they were linearized for backbone removal by enzyme restriction or PCR and used for yeast transformation. All DNA amplifications by PCR were done using Platinum SuperFi Master Mix (Thermo Fisher Scientific, Vilnius, Lithuania). All reactions were set up and performed according to the manufacturer's instructions for a high GC content template. All primers used in this study are available in **Supplementary Table S1**. All genes with their corresponding codon optimization for *R. toruloides* are available in **Supplementary Table S2**.

2.3. Yeast transformation and screening of transformants

R. toruloides was transformed according to the protocol described by Nora et al. (2019). To screen correct transformants, yeast colonies were checked by PCR and sequencing. The yeast colony was grown overnight in 10 ml of YPD and a quick genomic DNA extraction protocol was employed as described (Lõoke, Kristjuhan and Kristjuhan, 2017). Colonies considered positives were stored in 20% glycerol stocks at - 80 ˚C

freezer. Further verification was performed in terms of gene expression, by extracting the RNA, followed by cDNA synthesis, and its use as a template for the PCR and real-time PCR (RT-PCR) reactions.

2.4. RNA extraction and cDNA synthesis

Cell liquid culture from the mid-exponential phase was collected and harvested at 23,000 *g* for 30 s at 4 ˚C. After centrifugation, the supernatant was quickly discarded and the cell pellet was frozen in liquid nitrogen. Samples were stored at − 80 ˚C. RNA extraction was done using the Tri reagent kit (Invitrogen, USA), following the manufacturer's protocol. RNA pellets were resuspended in RNAse-free water and kept at − 80 ˚C. cDNA synthesis was performed with the First Strand cDNA Synthesis Kit (Thermo Fisher Scientific, Vilnius, Lithuania) according to the manufacturer's instructions for GC-rich templates.

2.5. Measurement of relative expression of genes by real-time PCR (RT-PCR)

Relative gene expression by RT-PCR was used to validate the overexpression of the native HMG1 genes. GAPDH (glyceraldehyde 3-phosphate dehydrogenase) was used as an endogenous control (Bonturi et al., 2017). Three biological and five technical replicates were used. The resulting cDNA from the previously described method was diluted 2.5 times before use. The reaction was done as follows: 2 µl of HOT FIREPol® EvaGreen® qPCR Supermix (EG) (SolisBiodyne, Estonia), 3 mmol/l of each forward and

reverse primer and 1 µl cDNA and Milli Q purified water to a final volume of 10 µl. Reaction conditions were: 95 ◦C for 12 min and 45 x (95 ˚C for 15 s, 64 ˚C for the 20 s, 72 ˚C for 20 s). Relative gene expression was calculated as described in Bonturi et al. (2017).

2.6. Growth in shake flasks

YPD media was used to characterize strains presented in **Table 1**, with 10, 20, 30 or 40 g/L of glucose, in flasks of 250 mL with 50 mL of medium. The cells were incubated at 30 ˚C at 200 rpm for 120 hours. For all strains, samples were taken for yeast characterization in terms of growth (OD600) and sugar consumption, measured by HPLC. A 15% (v/v) overlay of dodecane was added to the culture for monoterpene detection.

2.7. Growth in bioreactors

The strains SBY143 and SBY150, mutants containing the pinene cassettes, were grown in 1-liter bioreactors containing 800 mL of YPD media with 20 g/L of glucose. The culture was maintained at 30 ˚C and 200 rpm for 48 hours with pH at 6.5. Samples were taken for the yeast characterization in terms of growth (OD600), sugar consumption (measured by HPLC), and ethanol samples from the ethanol traps were collected for GC-MS analysis.

2.8. Analytical methods

Microbial growth was estimated by measuring OD600. Concentrations of metabolites, such as glucose, were measured using high-pressure liquid chromatography (HPLC) (LC-2050C, Shimadzu, Kyoto, Japan) equipped with HPX-87H column (Bio-Rad, CA, USA) and a refractive index detector (RID-20A, Shimadzu, Kyoto, Japan), at 45 ˚C and 5 mmol/l H_2SO_4 as mobile phase with isocratic elution at 0.6 ml/min. For monoterpene quantification using GC-MS, either the dodecane phase was extracted by pipetting from the cultures or the ethanol was collected from the ethanol traps and injected directly into the GC. The method used in the GC-MS was based on the one published by Kirby et al., 2021. Standard curves were made with α-Pinene (Sigma-Aldrich) diluted in both dodecane and ethanol ranging from 0.1 to 1 mg/L.

3. Results

3.1. Designing constructs for monoterpene production

To achieve the production of monoterpenes using metabolic engineering strategies, some constructs were designed for genetic engineering of *R. toruloides.* Monoterpene synthases are the enzymes that catalyze the chemical reaction of geranyl diphosphate (GPP) to monoterpenes and are generally found in plants. Some strategies were adopted to express and optimize the production of these monoterpenes, considering the presence of their precursor in cells. The first construct expresses the gene encoding geranyl diphosphate synthase (GPPS), an enzyme of the mevalonate pathway that produces the pinene GPP precursor, and pinene synthase (PS), both genes from the genome of the *Abies grandis* tree, as described in the work by Sarria et al. (2014). Sarria

et al. (2014), where they tested several pinene synthases from several plants and found the best one to be from *A. grandis*. The authors obtained better pinene production in *E. coli* when this PS was combined with the GPPS enzyme from the same tree species (Sarria et al., 2014). The second construct contains the tomato neryl diphosphate synthase (NDPS1) gene, which produces neryl diphosphate (NPP) as an alternative substrate to GPP, thus avoiding competition in the use of the substrate based on the work of Wei et al. (2021), in addition to the PS enzyme from *A. grandis*. The constructs containing the other two monoterpene synthases were based on the same strategies, replacing the genes encoding the PS enzyme with their respective genes. Geraniol synthase (GES) from basil (*Ocimum basilicum*) and linalool synthase (LIS) from mint (*Mentha citrata*). A schematic representation of the MVA pathway and genes that were heterologously expressed in this study can be found in **Figure 1**.

Moreover, we also chose a fourth monoterpene synthase based on the work from Kirby et al. (2021), where they successfully produced the monoterpene 1,8-cineole using heterologous enzymes. We synthesized the same sequence published in their paper to have one gene we already know that works for *R. toruloides*. In their work, they also express the enzyme GPPS from *A. grandis*. They created a truncated version of the protein without the signal peptide for the plastid and codon optimized for *R. toruloides*. Based on that, all monoterpene synthases in this work had their plastid signal peptides truncated to remove the contained in the native versions of their respective plants. All enzymes were truncated using the strategy adopted by Denby et al (2018), where they found a double arginine residue to be essential for the plastid signal peptide. So, all amino acids from the ATG until the double arginine residue were deleted.

A schematic representation of how the constructs were built can be seen in **Figure 2**. In total, there were seven combinations of constructs: GPPS-PS, NDPS1-PS, GPPS-HYP3, GPPS-LIS, NDPS1-LIS, GPPS-GES, and NDPS1-GES. The plasmids containing PS and HYP3 were built using the bleomycin resistance marker and the plasmids for GES and LIS were built using the nourseothricin resistance marker.

Regarding the other genetic parts that comprise the constructs, the promoters were chosen based on the constitutive promoter library created by Nora et al., (2019). The P9 promoter, the promoter for histones 3 and 4, is strong and bidirectional and, therefore, expresses the monoterpene synthases and the resistance gene simultaneously. P10 is the promoter for ubiquitin C and is a medium-strength one. We decided to use a less strong promoter to express GPPS to reduce negative feedback from the excess of substrate. The CAR2 homology regions were chosen so the construct replaced the phytoene synthase enzyme gene, which is part of the beta-carotene production pathway. This region was selected for two main reasons: first, to reduce the flux of GPP to a competing pathway and, secondly, because, in the haploid strain, this deletion facilitates the selection of positive mutants by making them albino (Nora et al., 2019).

All of the expression cassettes designed here were successfully assembled using the Golden Gate Assembly platform developed for *R. toruloides* (Bonturi et al., 2022), and confirmed by sequencing.

3.2. Screening of mutant stains

Two parental strains of *R. toruloides* were transformed with the developed expression cassettes: *R. toruloides* SBY29 and *R. toruloides* IFO0880. While the transformation for strain SBY29 worked using the expression cassette containing the bleomycin resistance gene, none of these constructs worked for strain IFO0880. New constructions of the same expression cassettes were developed, having resistance to nourseothricin, but the transformations again did not work for this strain.

The selection of yeast mutants containing the expression cassettes inserted into their genome was performed by colony PCR. Once colonies were found positive by PCR, they were grown in liquid culture for RNA extraction and cDNA synthesis. Another round of PCR was performed on the cDNA samples to verify that heterologous genes were properly expressed. Strains that showed heterologous genes integrated into the genome and expressed were considered real transformants and were included in the laboratory's strain registry. The positive transformants obtained were the SBY29 strain containing the GPPS-PS construct and the SBY29 strain containing the GPPS-HYP3 construct **(Figure 3)**. The same confirmation methodology was performed for the linalool and geraniol-producing transformants. However, the expression of GPPS and NDPS1 were not detected, only the expression of genes encoding linalool synthase and geraniol synthase **(Figure 4)**.

3.3. Overexpression of native HMG1 gene

Subsequently, we decided to use these same strains and overexpress the native HMG1 gene, which is also part of the mevalonate pathway, increasing the flow of the

pathway for the production of the heterologous terpene, an approach also used by other studies that produced terpenes in yeast (Kirby et al. al., 2021). One of the strategies was to use a strain from the laboratory repository that already contained the overexpression of the HMG1 enzyme (*R. toruloides* SBY29 with overexpression of HMG1) and transform the construct containing the gene encoding PS into this strain. This transformation was also confirmed by RNA extraction and PCR from the resulting cDNA **(Figure 5A)**.

Another strategy was to build an expression cassette containing the native yeast HMG1 and transform it into the GPPS-PS positive strain. Thus, a new positive transformant was obtained: *R. toruloides* SBY29 containing HMG1-GPPS-PS. Positive colonies of the transformants were grown in liquid culture for RNA extraction to verify the overexpression of HMG1. This overexpression was confirmed by real-time qPCR comparing native HMG1 expression with the expression of the same gene from the transformants **(Figure 5B)**. A summary of the positive transformants obtained can be seen in **Table 1**.

3.4. Detection of monoterpenes using dodecane overlay

The evaluation of the production of monoterpenes by the strains obtained needed to be carried out through gas chromatography coupled to mass spectrum (GC-MS), since these compounds are small and volatile molecules (Jiang, Kempinski and Chappell, 2016). Some tests were conducted to verify the presence of monoterpenes in the cultures, starting with the detection of pinenes. The first assay was performed by growing the transformed strains SBY143 and SBY144 in liquid culture in 250 mL flasks with 50 mL of

YPD medium with 10 g/L or 30 g/L of glucose. **Figure 6** shows the growth curves and sugar consumption of the cultures grown with 30 g/L of glucose. The cultures contained an overlay of dodecane 15%, where the pinene was to be trapped. After 48 hours of cultivation, the dodecane layer was removed and sent for analysis in the GC-MS. The GC-MS used for the first analysis showed peaks of dodecane remaining in the system even long after the injections with the samples. So, to avoid contaminating the GC system with dodecane, the injection amount had to be reduced, thus reducing the detection limit. Because of this, no significant amount of pinene was detected in these samples.

Due to these analytical difficulties in detecting pinene in the samples, we sent the mutant strains to our collaborators, Dr. Young-Kyong Park and Dr. Rodrigo Ledesma-Amaro at Imperial College London, who carried out a similar experiment. They cultured strains SBY143 and SBY150, and SBY29 as a control in YPD containing 40 g/L of glucose with 20% dodecane overlay. Dodecane overlay was removed after 120 hours of cultivation and added to the GC-MS for the detection of pinene, but no pinenes were detected in the samples **(Figure 7)**.

Then, we performed a similar strategy for detecting linalool and geraniol, growing the strains in YPD containing 40 g/L of glucose with 20% dodecane overlay. Flasks were incubated at 30 ˚C for 120 hours at 200 rpm. In **Figure 8**, the growth curves and glucose consumption of these strains are shown. Dodecane overlays from all the samples were then sent to collaborator Dr. Stéphanie Baumberger at the Institut Jean-Pierre Bourgin from AgroParisTech-Université in Versailles, France. There, they performed GC-MS analysis, but no detectable amounts of monoterpenes were found in the samples.

3.5. Detection of monoterpenes in ethanol

Since no monoterpenes were detected in the cultures, and several issues were arising with the use of dodecane in the CG-MS systems, we decided to pursue other solvents that could be less damaging to the equipment. Then, we developed a new assay in an attempt to detect monoterpenes in ethanol. To perform this, we carried out an experiment growing strains SBY143 and SBY144 in 1 L bioreactors **(Figure 9)**. Our attempt was to capture the pinenes in bottles coupled in series to the bioreactor condenser. The bottles contained 100% ethanol (which we called "traps"). The reasoning was that the monoterpenes would be more soluble in ethanol than in the culture medium. Tests were performed by growing the native strain and adding commercial pinene in the bioreactors, and up to 30% of the pinene was recovered in ethanol, as detected by GC-MS **(Table 2)**. However, when the mutant strains were used for cultures in the bioreactors, no detectable amount of pinene was found in any of the traps.

3.6. Overexpression strategy

Since no monoterpene production was detected in any of the mutants obtained, we decided to build new versions of the expression cassettes. This time, we decided to remove the CAR2 homology region to allow random integrations and, hopefully, have more integrations in the genome resulting from the same transformation. As shown in previous works with this yeast, more gene copies correlate with higher production of terpenes (Kirby et al., 2021). Additionally, several different variants were built with a wide

range of antibiotic-resistance genes so that we could do more rounds of transformations. We also added a heterologous and truncated form of the HMG1 enzyme gene from the genome of the yeast *Y. lipolytica*, aiming to increase the metabolic flux of the mevalonate pathway. The truncated enzyme exposes the C-terminal catalytic domain to the cytosol, preventing proteolytic degradation and improving catalytic activity (Li *et al.*, 2022). This gene is being modulated by a very strong promoter, P14, which is the promoter for elongation factor 1-alpha, also from Nora et al. (2019). A schematic representation of the new overexpression cassettes can be seen in **Figure 10**. The new cassettes were successfully built by Golden Gate and confirmed by sequencing. The plasmids were linearized for transformation in *R. toruloides*. Each specific monoterpene synthase construct was transformed into its respective mutant. Thus, the new PS constructs were transformed into SBY143, the new GES constructs were transformed into SBY158 and LIS constructs were transformed into SBY156 **(Table 1)**. Colonies were obtained for the transformation containing the Hygromycin resistance marker. For each construct, 6 colonies were selected for genomic extraction and PCR for the Hygromycin resistance gene and the respective MS gene. **Figure 11** shows the electrophoresis gel for screening the positive transformants. The selection for real positives is still ongoing; therefore, these strains were not named like the previous mutants. Additionally, detecting monoterpenes in these overexpressed transformants is still in progress.

3.7. Toxicity test of α-pinene in culture

As previously shown in the literature, pinenes are toxic to cells, whereas α-pinene is more toxic than β-pinene (Sarria et al., 2014; Wei et al., 2021). Therefore, in parallel with the metabolic engineering and analytical chemistry experiments, the characterization of the growth of *R. toruloides* strains growing in a medium containing different concentrations of pure α-pinene was carried out to verify the toxicity of this compound. Results of this test showed that, when added to the medium in concentrations of 1 g/L, 3 g/L, 7 g/L and 10 g/L, α-pinene did not affect growth, only extended the lag phase, since cells only started to grow after about 100 hours of cultivation (**Supplementary Figure S1A**). However, this could also have been caused by using a minimal medium with xylose for cultivation. Yet, growth was not affected when the yeast was grown in a YPD medium with 1 g/L of α-pinene. Only when α-pinene was diluted in ethanol before adding it to the media, the growth was impaired (**Supplementary Figure S1B**), which could also have been caused by ethanol itself. More tests would need to be carried out to confirm whether or not there is a toxic effect from the presence of α-pinene in the culture medium. However, as long as we are testing the production using the dodecane overlay to trap the monoterpenes, we avoid the toxic effect of having the compounds diffuse into the culture.

4. Discussion

Engineering microorganisms for the production of monoterpenes is still a challenge. One of the reasons is that monoterpenes are very toxic to the cells. And another one is that the engineering strategies must be well fine-tuned to achieve a robust, high-producing strain. Studies have demonstrated the production of monoterpenes in the model organisms, *E. coli* and *S. cerevisiae*, and even in the well-studied non-conventional

yeast *Y. lipolytica* (Oswald *et al.*, 2007; Sarria *et al.*, 2014; Pardo *et al.*, 2015; Niu *et al.*, 2018; Zhou *et al.*, 2020; Wei *et al.*, 2021; Li *et al.*, 2022). Still, for *R. toruloides*, three studies have described the production of monoterpenes so far. While limonene and 1,8-cineole titers were above 300 mg/L and 1.4 g/L, respectively (Kirby et al., 2021; Liu et al., 2021), Zhuang *et al.*, tried to express 16 monoterpene synthases genes, having succeeded with 8 of them, with their titers ranging from 5 to 20 mg/L for ocimene, carene, sabinene and pinene, thujene, 1,8-cineole and limonene. Nevertheless, in our study, the combination of synthetic biology and metabolic engineering strategies collaborated for a better understanding of the limitations to achieve such a bold goal, and successive cycles of DBTL cycle were applied to overcome these limitations.

A schematic representation of the biochemical pathway engineered in this work is depicted in **Figure 1**. The MVA pathway starts with Acetyl-CoA, which is condensed to form acetoacetyl-CoA. A second condensation forms 3-hydroxy-3- methyl-glutaryl-CoA and the reduction of this molecule results in mevalonate. Mevalonate is converted to IPP and DMAPP. GPPS then converts them to GPP, the substrate with 10 carbons (C10) used to form monoterpenes. Molecules further in the pathway (in **Figure 1** shown as "native isoprenoid pathway") can be used to form other isoprenoids, like FPP (C15), which can be used to produce molecules like sesquiterpenes and GGPP (C20), which is used by the cell to form the C40 chains of carotenoids (Wang, Quan and Xiao, 2019; Feng *et al.*, 2020). In **Figure 1**, we can also see which heterologous enzymes were expressed in our constructs to produce non-native monoterpenes. Even though we could not detect terpenes in our strains, we improved our knowledge of this pathway as a whole and can also give perspectives on how this pathway can be engineered even further.

One of the main takeaways from engineering this pathway is the need to downregulate competing pathways. By inserting our first construct in the CAR2 locus, disrupting the phytoene synthase gene, we were already aiming to reduce one of the competitors for substrates. However, since our transformations only worked for the SBY29 strain, which is multiploid, disrupting only one locus did not cause such an effect. Since the strain has other loci that still produce carotenoids, the competition over GPP was still occurring in the cell. In **Figure 7** we can see that strain SBY150, that overexpresses the native HMG1, is more orange, probably because the flux of substrates from the pathway is still being converted to carotenoids. To overcome that, other strategies of downregulation can be applied in the future. For instance, one that knocks down mRNAs instead of genes could work better for multiploid strains, like RNA interference (Liu *et al.*, 2019), or one that has multiple targets, like CRISPR (Otoupal *et al.*, 2019). These strategies, however, are still challenging and laborious for application in this organism.

Moreover, Zhang et al., (2019) have shown a preference of *R. toruloides* for converting FPP into sesquiterpenes than GPP into monoterpenes. FPP is synthesized by FPP synthases (FPPS), further down in the MVA pathway. To circumvent that, Kirby et al., 2021 presented a strategy to use an FPPS with mutations reducing the prenyl phosphate chain length elongation beyond C10 (GPP). Their most successful enzyme was the mutant FPPS (*GgFPS(N144W)* from *Gallus gallus.* This enzyme can also be applied in future engineering of the constructs.

On the other hand, an interesting alternative approach would be screening for monoterpenes synthases from other organisms. For example, the PS gene from *Abies*

grandis was chosen in this work. Still, in work by Wei et al. (2021), using *Y. lipolytica* as a host, they expressed the PS gene from *Pinus taeda*, achieving high yields of pinene production: up to 33.8 and 36.1 mg/L when using waste cooking oil and lignocellulosic hydrolysate as media, respectively.

Furthermore, here, we also managed to use the constitutive promoters - that were only tested with fluorescent proteins so far - in real-life applications, modulating enzymes from biochemical processes. We showed that promoter P9 (the promoter for histones 3 and 4), predicted to be bidirectional, presented a strong expression of the two genes flanking it, as shown in the cDNA expression assays (**Figures 3 and 4**). The strong expression is also true for the tHMG1 gene modulated by promoter 14. Contrastingly, promoter P10, the promoter for ubiquitin C, showed a less reliable behavior, since it did not work for most cases, as shown in **Figure 4**. Thus, we discourage future researchers from using this promoter when looking for a medium-strength promoter for their constructs. Nevertheless, the constructs built in this work can now be further developed for successful heterologous expression in future endeavors.

5. Conclusion

This worked paved the way for the expression of monoterpene synthases genes in *R. toruloides*. Several iterations of the DBTL cycle were followed in the pursuit of a successful metabolic engineering strategy for heterologous expression in this microbial host. Still, the strains obtained through the overexpression strategy still need to be tested with novel attempts of GC-MS for detection of the target monoterpenes. The ultimate goal is to grow the mutant strains in wood hydrolysates and obtain significant amounts of

monoterpenes, which are competitive with what has already been published in other organisms. Thus, confirming the role of *R. toruloides* as a sustainable microbial cell factory in the circular bioeconomy.

6. Figures

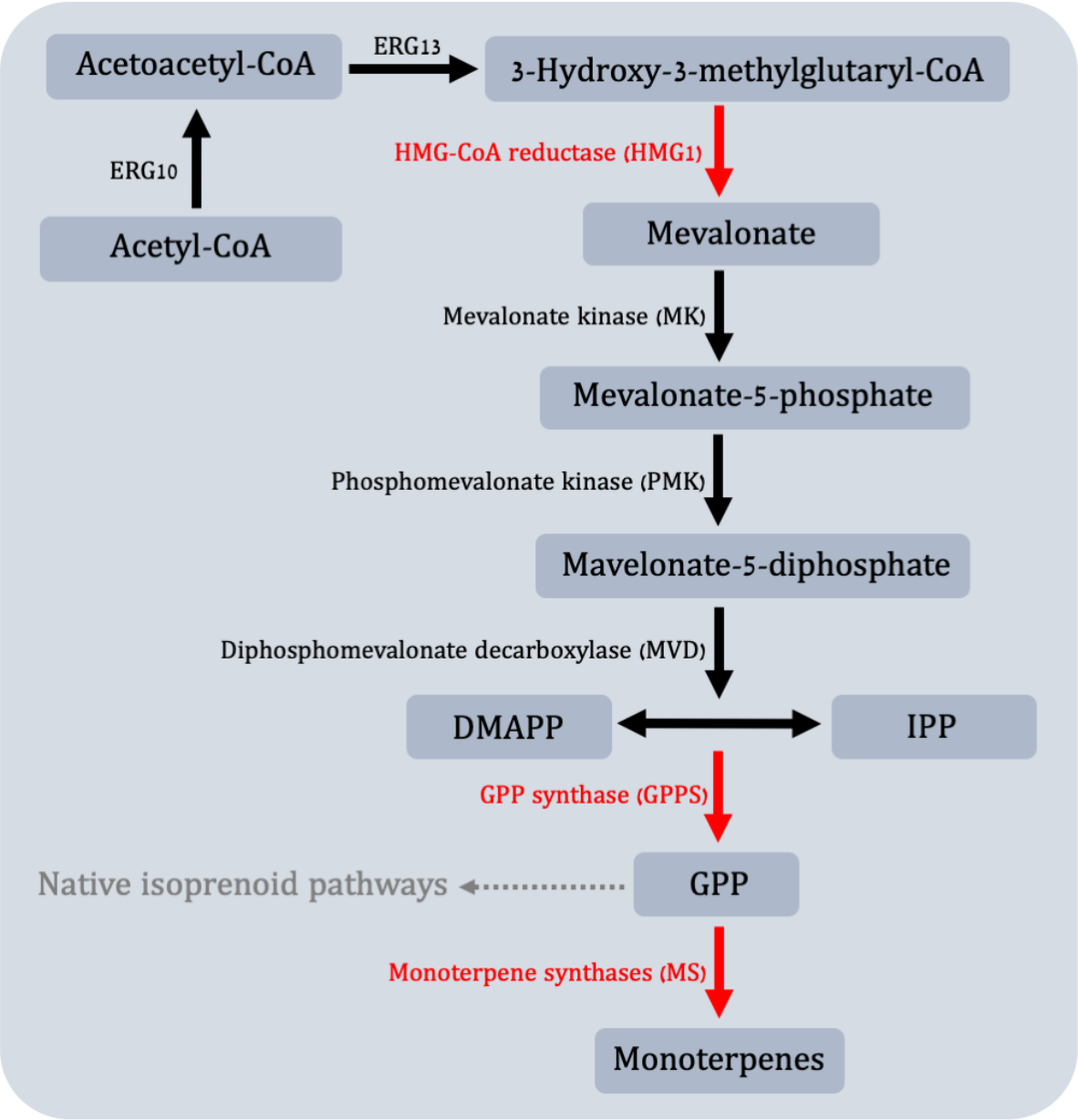


Figure 1. Schematic representation of the mevalonate (MVA) pathway. Red arrows represent the enzymes that had their heterologous expression tested in this work.

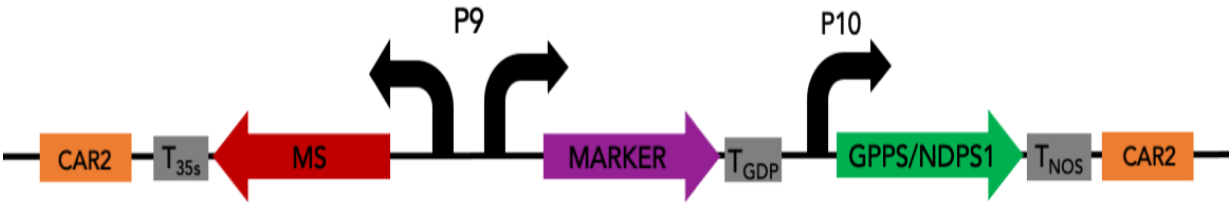


Figure 2. Scheme representing the expression cassettes for monoterpene production. CAR2 are the homology regions to the phytoene synthase-producing gene. MS are the monoterpene synthase genes. P9 is a bidirectional promoter modulating both the MS genes and the markers. Marker is the resistance marker that can be either NAT or BLE. P10 is a medium-strength promoter modulating either the GPPS enzyme gene or NDPS1. The terminators used for each gene were t35s, tGPD and tNOS.

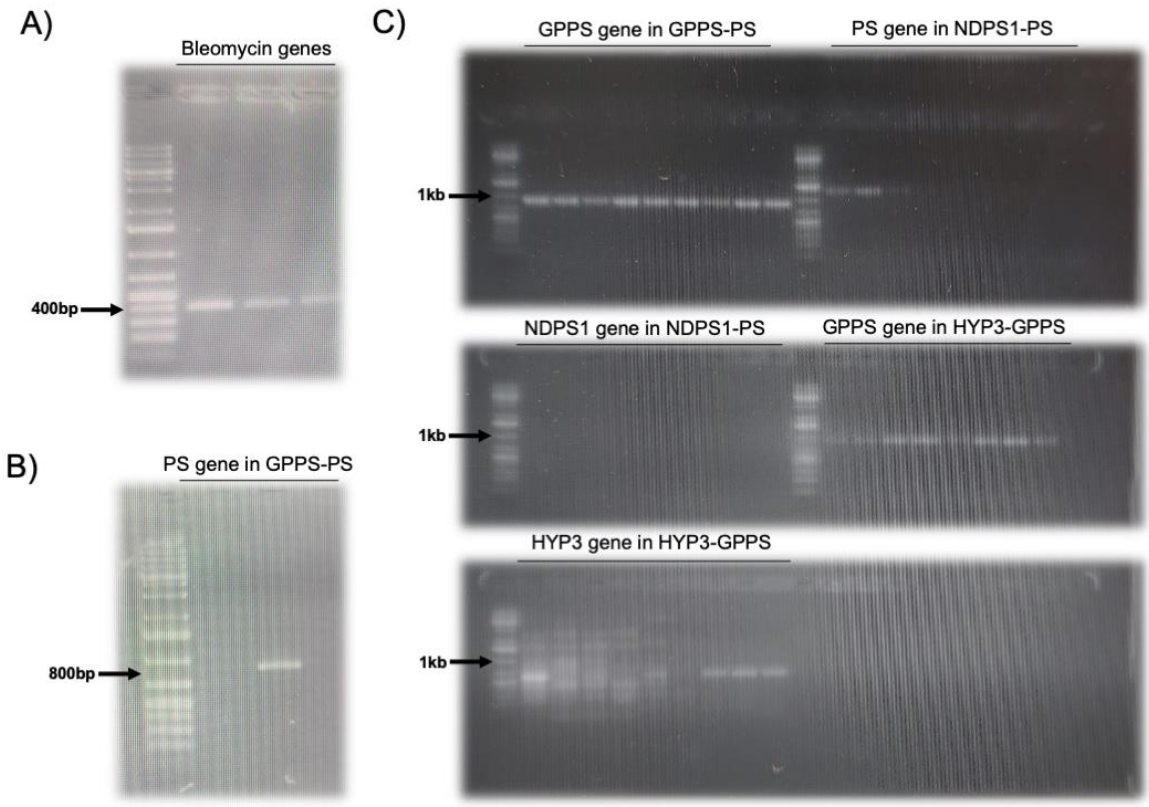


Figure 3. Electrophoresis gel with PCR reactions using cDNA in a temperature gradient to test gene expression in candidate transformants. **A)** Bands of 400bp referring to the expression of the gene that confers resistance to Bleomycin. **B)** Bands of 800bp referring to the expression of the gene coding for the PS enzyme. **C)** Bands referring to the expression of the genes of each construct as described in the image, 1kb marker was highlighted to facilitate viewing. The molecular marker used was 1kB Plus from Thermo Fisher Scientific (Vilnius, Lithuania).

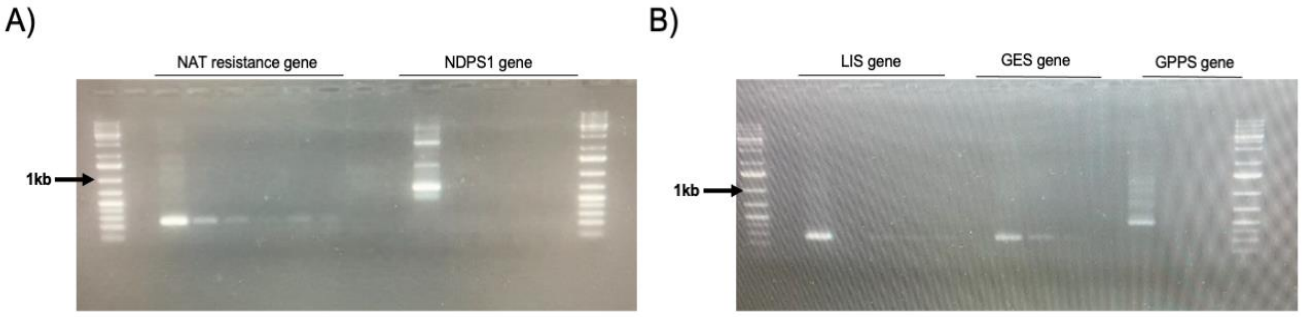


Figure 4. Electrophoresis gel with PCR reactions using cDNA in a temperature gradient to test gene expression in candidate transformants. **A)** Expression of the gene that confers resistance to nourseothricin and the gene that encodes the NDPS1 enzyme. **B)** Expression of genes coding for LIS, GES and GPPS enzymes. The first band of each gene is the positive control reaction using the plasmid as a template. The molecular marker used was 1kB Plus from Thermo Fisher Scientific (Vilnius, Lithuania).

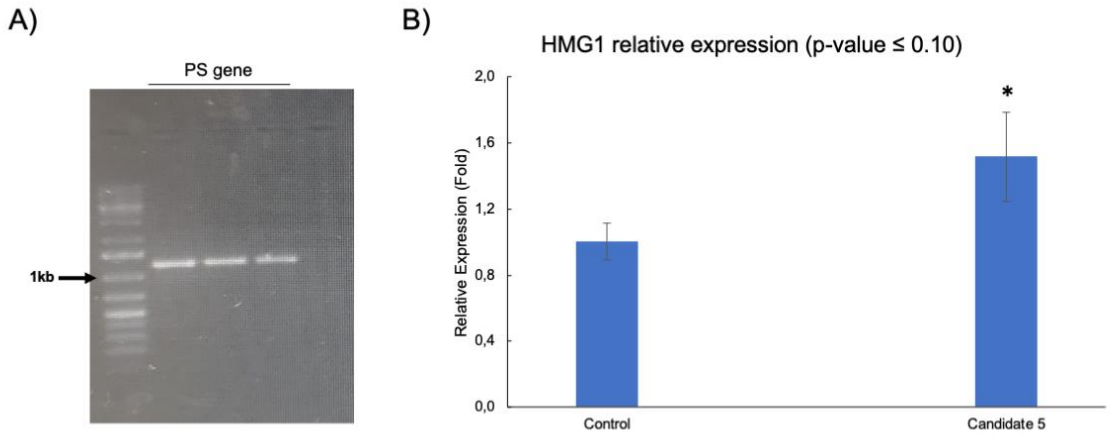


Figure 5. Confirmation of strains overexpressing HMG1. **A)** PCR reaction using cDNA was performed to confirm the pinene synthase enzyme gene expression in the SBY92 strain containing HMG1 overexpression. **B)** qPCR to confirm the overexpression of the gene encoding HMG1 in two candidate transformants compared to the native yeast gene without any modification. Six candidates were tested in total, but based on the T-test with a p-value <0.1, candidate 5 demonstrated a statistically relevant increase in relative expression and was chosen as the positive mutant strain for this transformation. The other candidates presented a relative expression value lower than the control and are not shown in the graph for clarity.

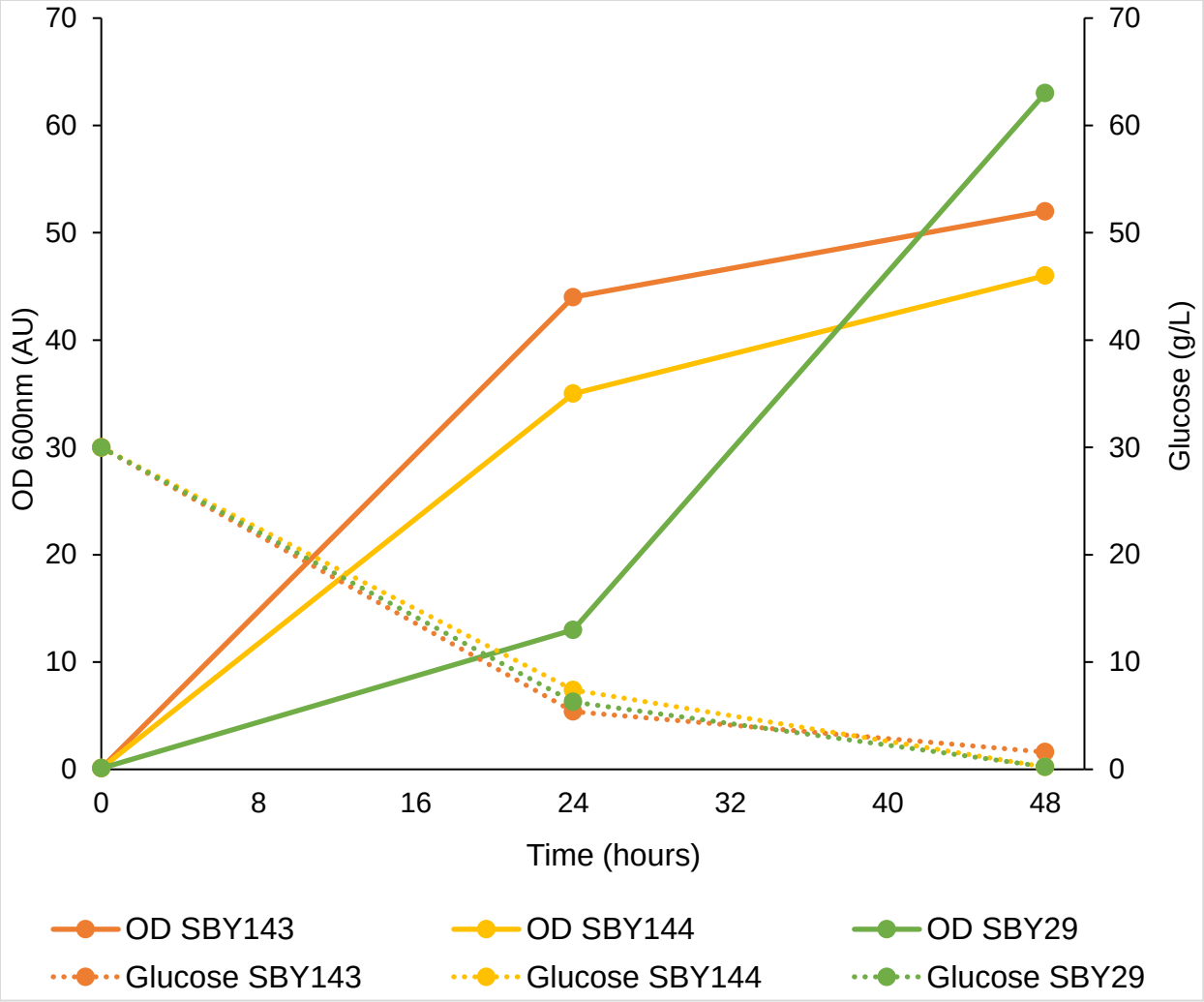


Figure 6. Culture of mutant strains containing pinene constructs (SBY143 and SBY144) and parental strain SBY29 grown in YPD with 30 g/L glucose. The dodecane layer was extracted at 48 hours for analysis by GC-MS.

Figure 7. Culture of strains for GC-MS analysis. Cultures were grown in YPD containing 40 g/L of glucose (10 mL in 50 mL shake flasks) with 20% dodecane layer. Flasks were parafilm-sealed and were incubated at 30 ˚C for 120 hours at 200 rpm and dodecane layer was removed at the end of the culture for GC-MS analysis. This experiment was done in duplicate. Initial OD of all cultures was 0.1 and final OD for each strain was: SBY29 95.90 ± 8.10; SBY143 81.35 ± 11.5; SBY150 125.85 ± 12.45.

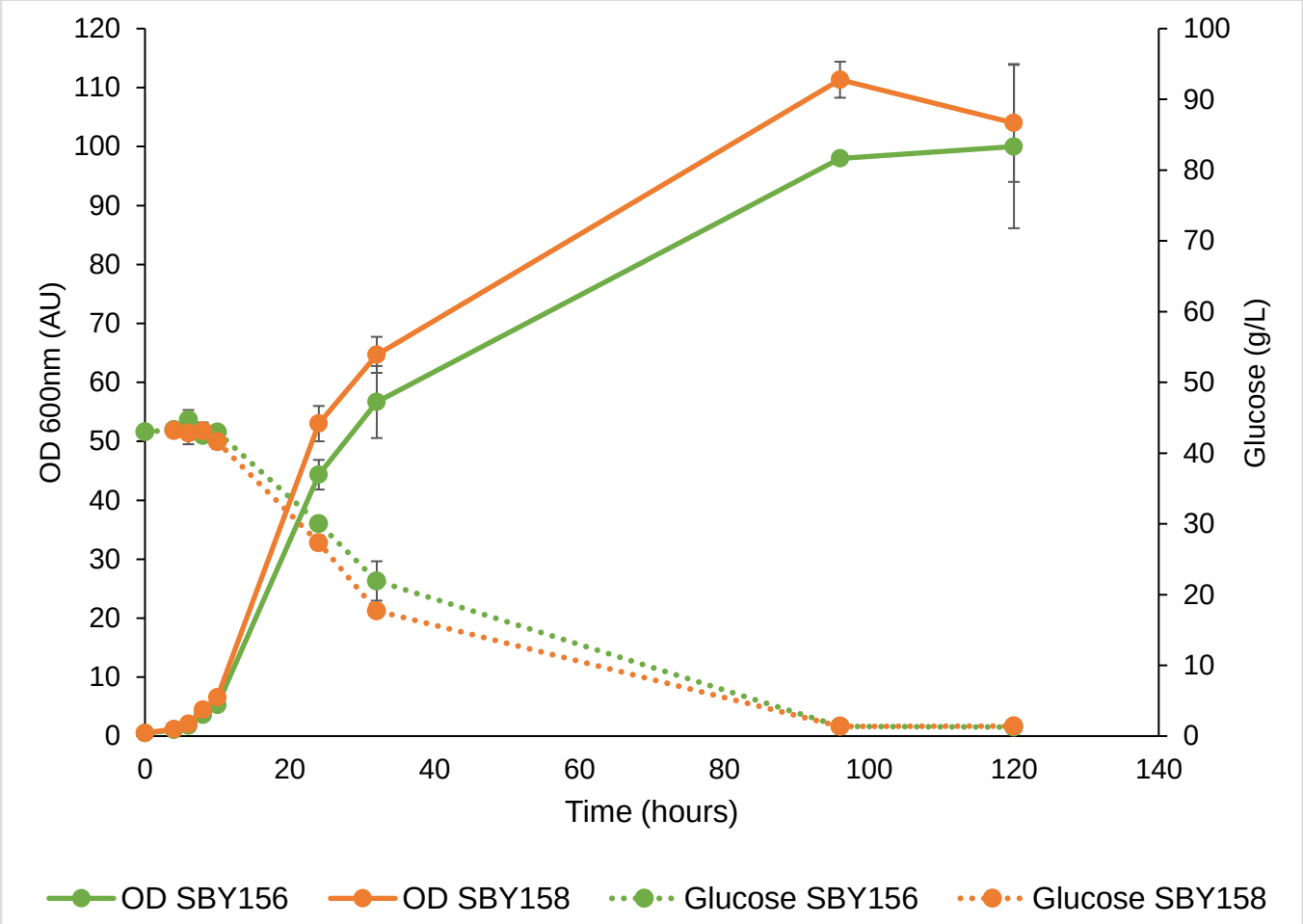


Figure 8. Culture of mutant strains containing linalool (SBY156) and geraniol (SBY158) constructs grown in YPD with 40 g/L glucose. The dodecane layer was extracted at 120 hours for analysis by GC-MS.

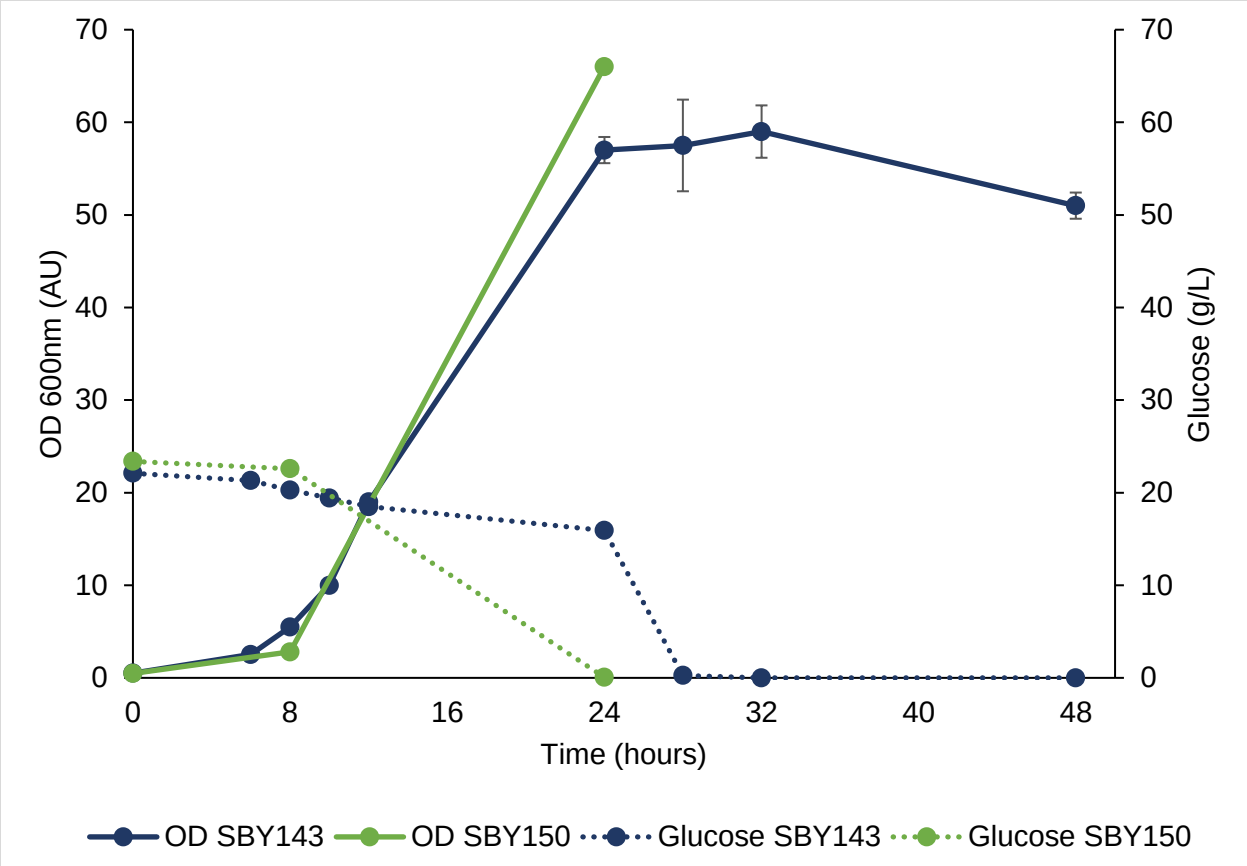


Figure 9. Growth of SBY143 and SBY150 strains in bioreactors with ethanol traps. The SBY143 experiment was run in two bioreactors; the average and standard deviation from the results of the duplicates is shown. The SBY150 experiment was run in one bioreactor. All experiments were carried out in YPD with 20 g/L of glucose. Ethanol samples were collected from the ethanol trap at 11, 23, 28, 32 and 48 hours of cultivation and sent for analysis in the GC-MS.

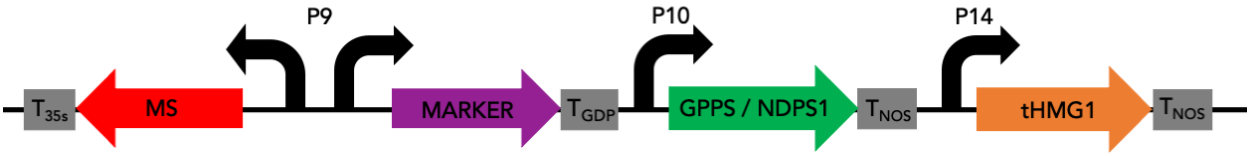


Figure 10. Scheme representing the overexpression cassettes for monoterpene production. The homology region was removed to allow for random integrations. MS are the monoterpene synthase genes. P9 is a bidirectional promoter modulating both the MS genes and the markers. Marker is the resistance marker that can be either NAT, BLE, HYG or G418. P10 is a medium-strength promoter modulating either the GPPS enzyme gene or NDPS1. P14 is a strong promoter modulating the expression of a truncated HMG1 gene.

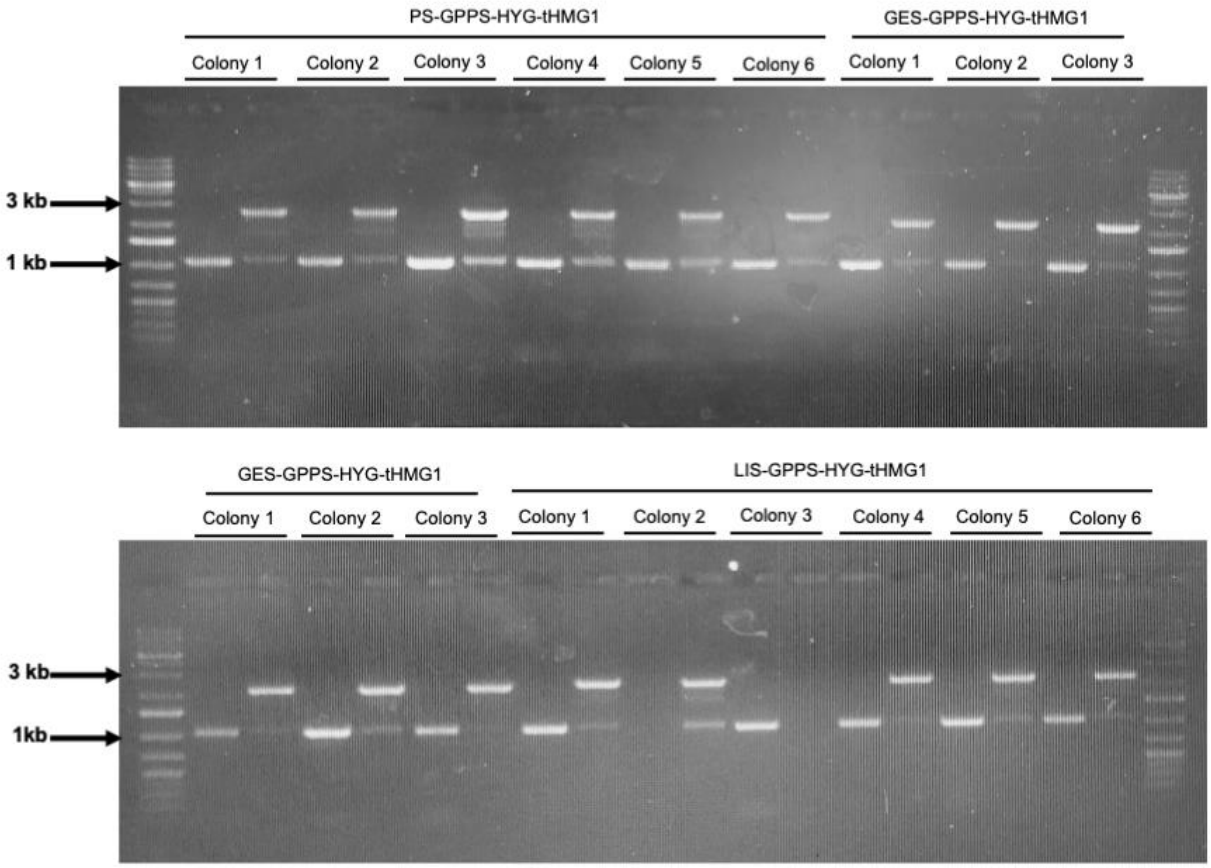


Figure 11. Electrophoresis gel for screening of positive transformants with the overexpression cassette. The gene for each monoterpene synthase was amplified (bands around 2kb), and the gene for hygromycin resistance (bands around 1kb). The molecular marker used was 1kB Plus from Thermo Fisher Scientific (Vilnius, Lithuania).

7. Tables

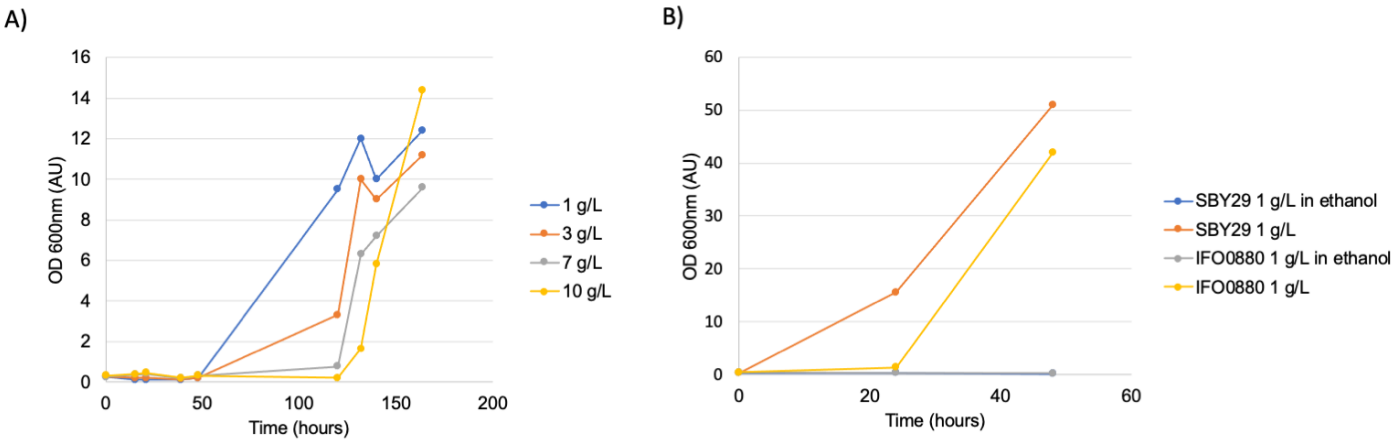
Table 1. Strains transformed with monoterpene synthase genes obtained and confirmed by PCR.

Strain	Background	Construct	Product	Confirmation	Resistance
SBY143	SBY29	GPPS-PS	Pinene	cDNA PCR	Ble
SBY144	SBY29	GPPS-HYP3	1,8-cineole	cDNA PCR	Ble
SBY150	SBY92	HMG1-PS	Pinene	cDNA PCR	Ble
SBY151	SBY144	HMG1-GPPS-PS	Pinene	qPCR	Ble
SBY155	SBY29	McLis-NDPS1	Linalool	cDNA PCR	NAT
SBY156	SBY29	McLis-GPPS	Linalool	cDNA PCR	NAT
SBY157	SBY29	ObGes-NDPS1	Geraniol	cDNA PCR	NAT
SBY158	SBY29	ObGes-GPPS	Geraniol	cDNA PCR	NAT

Table 2. Calculation of recovery of pinenes from traps with ethanol in cultures in bioreactors adding purified alpha-pinene.

	Replicate 1			Replicate 2			Replicate 3		
	Trap 1	Trap 2	Trap 3	Trap 1	Trap 2	Trap 3	Trap 1	Trap 2	Trap 3
Concentration (mg/L)	11,84	1,73	0,00	42,27	7,06	0,81	40,94	6,72	0,72
Final volume (L)	0,090	0,096	0,097	0,090	0,095	0,095	0,090	0,096	0,097
Pinene obtained (mg)	1,07	0,17	0,00	3,80	0,67	0,08	3,68	0,64	0,07
Pinene added (mg)	15,00	15,00	15,00	15,00	15,00	15,00	15,00	15,00	15,00
Recovery (%)	7,10	1,10	0,00	25,36	4,47	0,51	24,56	4,30	0,46
Total Recovery (%)	**8,2**			**30,3**			**29,3**		

8. Supplementary Material



Supplementary Figure S1. Pinene toxicity tests. A) *R. toruloides* IFO0880 grown in DELFT minimal medium containing 30 g/L xylose. Pinene concentrations tested were: 1g/L, 3g/L, 7g/L and 10g/L. **B)** *R. toruloides* IFO0880 and SBY29 grown in YPD with 1 g/L of purified pinene or 1 g/L diluted in ethanol.

Supplementary Table S1. Primers used in this study. BsaI cut sites are underlined and overhangs used for Golden Gate are in bold.

Name	Sequence 5'- 3'	Template	Application
P_CAR2F_F	ACGGGGTCTCG**ggag**ggctgtctgtgacttgctatc	CAR2 locus	Amplify upstream CAR2 locus to build cassette with GGA
B_CAR2F_R	ACGGGGTCTCG**acct**gacagcccagcggggc	CAR2 locus	Amplify upstream CAR2 locus to build cassette with GGA
B_t35s_rv_F	ACGGGGTCTCG**aggt**tctggattttagtactggattttgg	T35s	Amplify t35s to build cassette with GGA
T35s_rv_R	ACGGGGTCTCG**gacg**acgctgaaatcaccagtctc	T35s	Amplify t35s to build cassette with GGA
F_CAR2R_F	ACGGGGTCTCG**gctt**ctggtttctcaacttcgagcc	CAR2 locus	Amplify downstream CAR2 locus to build cassette with GGA
Q_CAR2R_R	ACGGGGTCTCG**atgg**tcgagctcaagctgaagaag	CAR2 locus	Amplify downstream CAR2 locus to build cassette with GGA
S_P9_F	ACGGGGTCTCG**cgca**gttctgtaggagagggttggtagg	P9	Amplify P9 to build cassette with GGA
T_P9_R	ACGGGGTCTCA**gttg**ctgttgcgggaagtgcgg	P9	Amplify P9 to build cassette with GGA
U_BLE_F	GCATGGTCTCC**caac**atggccaagcttacctcggc	Bleomycin gene	Amplify Bleomycin resistance gene to build cassette with GGA
tGPD:R_marker_R	TCTCTGGGTCTCC**acac**ttagactcccgcccctctcgc	tGPD	Amplify tGPD to build cassette with GGA
R_P10_F	ACGGGGTCTCG**gtgt**acggactgagcggtcgaag	P10	Amplify P10 to build cassette with GGA

D_P10_R	ACGGGGTCTCG**catt**gtctgtagtgagagcgaaatgagg	P10	Amplify P10 to build cassette with GGA
F_T1_tNos_F	GCATGGTCTCa**tcta**cgttcaaacatttggcaataaagtttc	tNos	Amplify tNos to build cassette with GGA
F_T1_tNos_R	ATGCGGTCTCa**aagc**cccgatctagtaacatagatgaca	tNos	Amplify tNos to build cassette with GGA
SQP_5(PS)	gcatatacgagtcaatgtaggcc	PS gene	Amplify gene to test cDNA expression
tAgPs_R1	ccttccatcagcatgagc	PS gene	Amplify gene to test cDNA expression
SQP_11(GPPS)	catggactctaaggctatgacgg	GPPS gene	Amplify gene to test cDNA expression
tAgGPPS_R1	gcggaaggcgacgtaatc	GPPS gene	Amplify gene to test cDNA expression
SQP_11(NDSP1)	gactgagaagctctgctacgagg	NDPS1 gene	Amplify gene to test cDNA expression
NDPS1_R1	cgaagcgcctgtggcg	NDPS1 gene	Amplify gene to test cDNA expression
SQP_5(HYP3)	ggccttgcgagtgtgctcaacc	HYP3 gene	Amplify gene to test cDNA expression
HYP3_R1	ccatcacgtgctcgttcgac	HYP3 gene	Amplify gene to test cDNA expression
gga-ps-fw	ACGGGGTCTCG**ggag**tctggattttagtactgg	GGA-PS	Amplify fragment without the homology regions
gga-ps-rv	ACGGGGTCTCG**acac**gttgagaaaccagaagcc	GGA-PS	Amplify fragment without the homology regions
tHMG1-fw	ACGGGGTCTCG**gtgt**cgcgaagcggtag	tHMG plasmid	Amplify P14-tHMG1-TNos for new overexpression cassette

tHMG1-rv	ACGGGGTCTCG**atgg**CACGTTCAGCAAtgacc	tHMG plasmid	Amplify P14-tHMG1-TNos cassette for new overexpression cassette
Tgpd_Fw	GCATGGTCTCa**tcta**cgagctccccggatcc	GGA-PS-BLE	Amplify cassette starting from from Tgpd
T_P9_R	ACGGGGTCTCA**gttg**ctgttgcgggaagtgcgg	GGA-PS-BLE	Amplify P9 to change the marker
hyg_fw	ACGGGGTCTCA**caac**atgaagAAACCTgagC	HYG plasmid	Change to HYG marker
hyg_rv	GCATGGTCTCa**taga**cgctactcTTTAGCgc	HYG plasmid	Change to HYG marker
g418_fw	ACGGGGTCTCA**caac**atgggcaaggagaagacc	G418 plasmid	Change to G418 marker
g418_rv	GCATGGTCTCa**taga**cctagaagaactcgtcgagc	G418 plasmid	Change to G418 marker
U_NAT_F	ACAGGGTCTCC**caac**caagccacgatggcggc	NAT gene	Change to NAT marker

Supplementary Table S2. Heterologous genes synthesized in this study. All genes had their codons optimized for *R. toruloides* and present truncation in the region for the plastid signal peptide.

Gene name	Sequence 5'- 3'	Source	Reference
Pinene Synthase	GGGTCTCACGTCCTACAAAGGGACAGATTCAAGGAGG GTGCGCGTGACGAGCGACTTCGTTTCGACGTTGGCGA CGCTGTAGCCGTCGCGGTACTTGTAGCCGTAGTGGAA AGCCCGGGCGATGTCGAACGCATGCTTCTTCGCCGAG ATGGGAACGTTGATGTCTGGCTTGAGAAGCTCCCAGT TGAGTCCCTTGATGACGTCACTGATCATGGCGTTGATA TGATCGAGGGCGTCTTCCTCGCTGACACCGGGGTTGT CCTTCATGTAGCATGAAATAGAGGAGGCTTCCTCTCCG CGAGCGCGGTCCGCCTTGTAGCACCGCGTATCACCTC GCAATCGAAGGATGGCGCATGCCAAGTCGTTAAGCTT CGAGGGGAAGTCGACCTCCTTGAGGATGTGGTCAGG GAAGGGGATGTCCATCGTCAGAATGGGCTGCAATGCG GAGATGCGGTGGCCACAGCTGACCTTCCCGTTCTCGT AGTACTCGTCGAACGAGGGCAGGTAGCCAGTGGCGAT CCAGCGTGCCTCTTGCATATACGAGTCAATGTAGGCCT CCCACGCTTCCCGCGCGTATGTGAGCGTGTCGCGGC CCTGCGCCTCCTCGGCCTCGCGAGCCATCTCATTGAC GGTGTCGTAGACCGCGATGTATACTCCCTTCATGTACT CTGGGAGGCAGTCGATCGAGGACGGATCCCACCTCTT CATGGTCGCGGTGAAGAGCTCGAGCTCGTCGACGGT GCCGAAGGTGTCGTACATGTCGTCGAGGACCGTGATA AGGTGGCACGTCTTGGCAAAGCCGAGGCGGAATCCA GAGTGTTGCGGCTCGAACGCGATGCAGGAGGCGAGG GTGTAGTACTCCACGTGGCGGTGGCGGCAGAAGGTCA TCTCAGGAAAGCCCGACTCCTTCCACCAGCGGACGAG CGATTCGAGCTCCCTCTTCTGGAGGGATTGGAAGATG TTGAACTCGAGTTTTGCGAGCTCGAGGAGCTTCTTGCT CTTGACGTACGACTTCGTGTTCTCCGTGTCCTGTCCAA AGACCTGGATGTAGTTGCGCGCTTCCAATCGCGGCAA ATACGTGTGCCAACCGTATTCGAGGACGTCCCCGATC TCGCGCGAGAGAGACGACACCGGAATCTTCTGCAGGG CCTCCTTGAGGTATTTGGTCGAGAAGATCTCAGCCTCG TCCATGATCTTCTCGCCTGGGAAGGCGATGAGGGAGG CCCGGAACAGATTGAGGACGCCTCTGATCTCCTCGTC TGTCTGAATGTTTTCCGAGCAGGAAAACTGGCCGTTCT GGCCCTTGAAAGCCTTGAAGACGTCCGACGACACCGG GTAGCCGTGGAGGCGGAGGGTGCGGAGGCCCAACGC AGTTGAGTTCAGGTCAGTAACGACCGACTCCCTCCCG CAGCCGATGCCGTTCTCGCCCCAGTACGAGTAAACGT	*Abies grandis*	Sarria et al., 2014

	AGTCGAGCGCCGACTTGATCTCGTCCTTGAAGTGGCG GTGGATGCCGAGGCGCTCGAGGCTGTCGACAATCCAA AGGCGCTGGATGAGGTCGTTGAGCGGCGACATGAGC TCGCCGTCCTCGAGCGACATCGAGTTGAACATGTTCTT GACCTCGCCGATGAGCTTCTCAGCACGCTCGAGGTAC GACTTTTCCTCGTACGCCGTCGGGAGCGACTGGATGA CGTCGTCGTCCCAGAGGTTGGAGTGGAAGTCGCCCAT GCGGCGGCGGACACCGTCGTCGGTGACGACGGTGGT CGAGCTCATGCTGATGGAAGGGGTGATCGACTTGCCT CGCCGCATCGCACGAGACCC		
Linalool Synthase	acgggtctcacgtcCTAGACATAGGGCTTGAACAGGAGGTT CGCGATCCGGTGCTGGAGCTGGGAGTGGTTGCCGTC TCCGTCGAGGTACATAAACTGCGCGTCGCGGCCAAGG TTAGCGGCCGCCATGACGACGTCACCGCGGAGGTCA GAGTTCGCCGCCATCGTGGTGTTCATCTCCTTCCACG CCTCGCGGATCATGAAGCGGACGTGCTCTTCGGCTTC TTCCCGCGTTGCATTCTGTTCCTTCATGTAAATCTGGA TCGACTTCGCGACGTCGCCTCTCTTCATCTCGAACGG GAGGGTGCCGAGGTCATCGGGAAGCCTGACGAGCAT TCCCGAAAGGCGGATGATCTCGTCGTACTCGTAGAAA CTCTCGATCACCGGCTTCTCCTTCGACTTGGCGAGCAT GAAGTAGATCTGGGAGATTACTGCCGGAGCGCCAATG GTGATGCTGGCGTAGTTGAGGTATTCCTCCATGGTGG GGGTGTACTTCGTGTAGTACCACTTTGCCTCTTGCAGG TAGCCTTCAACGAGGTCGACCCACGCACGGTGGAGAA ATGGGATGCTGATGAAGCCCTCCTCCTTGAGGATGTG GTACGCGACCTCGGAGACGAAGTTGTAGAGGACGAAG TAGAAGAGCTGCAAGTAGTACGGAAGCTGGGTGGCGG ACTCCGTATCCCACCGTCGGATGACGTGGGTGAAGAG CTGGAGCTCGTCGAGGGTGCCGTAGATGTCGTAGACG TCGTCGAGAGCTGTAATGAGGGTGATGATCTTTGCCG CCGTCTTGCGTTCGTAGCCGAACTTGTGAGCCTCAAA GAGGCCGAGGGCCCAATAGTAGCTCTCGACCAGCCTG TCGCGGACGAATGGGAGCTTCTGGGGAAGGCTCGAAT CGTTCCACCAGCGCGAGAGGTCCTTGAGCTCTTCCTG CTGCGTGGCTTGAATGATGTTGAAGTCGAGCTTGGCG AGCTCGAAGATGATAGGGTTCATGTCCTTGCGGCGTG AGTACGCGTCGAGGAACCAGCGCTCCTCCAAGCGTTG GATGCGCCAGTGGAGAGGGATCTCCAACGAGTGGCC GATCGACGACGAGAGGTTGTCACCGTCGATGCCGGCC TTGAGTTTCTCCTCGAGGAATTCGGTGGAGAACTTGCG AGCGACCTCGAGGGTGTCCTCTCCCTCGCGGAGAAG GAACGAGGCCTCGTAGAGTTGCAGCACTCCCTTCGTG TCTTCGCCGATGAGGGTTTTCTCGAAGTCCGAACCGT	*Mentha citrata*	

	GCTCGTTCTTGAAGCAGTCAAAGACGCCCTGCGAAAC GTTGAAGCCATGCTGCCGGAGGAGGCGGAAGCCAAG CGACGTGAAATGGAGGTCCCCGACCTTAGTCGAGTTA TTGTTCTGGAAAATCTTGTAGATAGAGTTCAAGATGTTC TTAATCTCGTCTCGGAAGAAGTACGACAATCCCAAGTT CTGCAAGTCGTCGATGAGCTCCAACTGCTTCGTGGCC TCCATCTTGGCCCCGAGGAGCTTCTTGACCTCCATGAT CAGCTCCTCCTCGCGCTCGAGCTGCTTCTCCTCCTTGT AGTGGTCAGTGTCGAGGCTCTGGATGAAGTCAAAGTC CCAGACCGAAGGGTGGTAGTTGCCGGAGCGTCGCATc gcacgagaccccgt		
Geraniol Synthase	ACGGGGTCTCACGTCCTACTGCGTGAAGAAGAGGGCG TCGACGTAGTTGTCGACGCTTGAAAAGTAAGTGTCCTG GTCGTGCTTGTAGACGACCTGAGACGCCCGCGCCATG TTAAGGGCGACCTTGATGATGGAGAGCGGGAGGTTCT TGTTGTAGACCAGCTCGCCGTTCAGGTCCCTCCAGAG GCCCTTGATCTCCTCGAGGATGCGACTGCGGGCCTCC TCCTCCGTGAGCGACTTCTCCTTCATGAAGAGCTGCAC GCACGAAGCCAGATCGCCGCGCTCCTGCTCCTCCTTG GCGGTTCCGAGGTCGTCCCAGAGGCGAAGAATGCGG CCGGCGGCGGAGAAGACCTTGGGGTAGGGTTTTTGG GTGAAGAGTTGGGAGTTCTGGTGTGTGACGCCTTCGC CGATGAGAAAGAAGATGTGTGCGAAAGCCATGTAGGC GCCTGCCGTGGAGACGCCGTTCTCGATATACTCCTCG AGCTTGGGTGCCGAGCCACCGTTGAACCACTTTGCTT CCTCCATGAAGCCTTCAATCATGTCGATCCACGTCGAC TTGAGGTTAAGGAGGACAATCCGGCCCGTGTCGCGGA GGACTTTGTAGCATACTTCATTGGTGGTATTGTACAAC GCCATGTAGCAGATCTTCATGTACTCAGGGAGGCCCT CCATCGCCTCAAGGTCCCACCTGCGGATGGCGTCGGT GAAGAGGATGAGGTCGTCCATCTCGCCGTAGGTGTCG AAGATGTCGTCGATGACGAGGAGAATAGAGATGGCCT TCGCGAGCTCAATGCGAACCGACGAGTACTTGGGCTC TGGGAGGAGGCCGACGGTCCAGAGGAAGCACTCGAG TGGTCTGTCTCGCCCAAACGAGAGCTTGTCGACAAGA CCGAGCTCCTTCCACCACCTGATGATCTCAGTGAGCT CGGACTGGTGCTGAGCCTGGACTTGGTTGTAGTCCAA GATGGCGAGCTCCAAAAGATCTCCGTCGTGGTCGCTC TGCTTGCCGTACTGCTCGATGAATCGGCGCGCCTCGA GGCGCGCCATGCGGAGATGGCGCGGGACGTCGAGCG CCTGCGCCACCTCACCGTGAAGAGGTGCTGCAGGTTC CGAAAGCGAACGACGAAGACGAGCTTCAGCGAACTCC ATAGCCTCCTCGAGGATCTCTTCGCCGGCAACTCCCA AGTTCGACGCTTCGTACAAGCTGAGGAGCCCCAGCGT GTCCGACTCGTCGAACTTTCCGCGCTCGTCCTTGAACT	*Omicum basilicum*	

	TGAGGAAGATTTCAGGGCTGATCTCGATGCCGTTGTG GCGGAGCAAGCGGAAGCGGAGAGCGGCGGTGAAGAG GTCTTCCTCTCCGGTCGAGAAAGGCGAGCGGAGGAC GGCGTTGATGGCGTCCTCAAAGTAGTAGCCGATGCCC AACTGCTGGATGTTGTCGATAAGCTTGAGCTTCTCGAC CGATTCGGTGTCGTTACGCTGGAGCTTTCGCGTCGTTT CCTCGAGCAGGTACTCCCGGCGCATCGCACGAGACC CCGT		
GPPS	GGGTCTCGAATGGTCGAGTTCGACTTTAATAAGTACAT GGACTCTAAGGCTATGACGGTCAACGAGGCGCTCAAC AAGGCGATCCCACTCCGCTACCCCCAGAAGATTTACG AGAGCATGCGGTACTCGCTCCTCGCAGGGGGCAAGC GCGTCCGACCGGTCCTCTGCATCGCCGCCTGTGAGCT GGTCGGTGGCACCGAGGAGCTTGCGATCCCCACCGC GTGCGCCATCGAGATGATCCACACGATGTCTCTTATGC ACGACGACCTCCCTTGCATCGACAACGACGACCTCAG GCGGGGAAAGCCCACCAACCACAAGATCTTCGGCGAG GACACCGCCGTTACGGCGGGGAACGCTCTCCACTCCT ACGCCTTCGAGCATATCGCAGTCTCGACGTCCAAGAC TGTCGGAGCGGACCGTATCCTCCGAATGGTCAGCGAG CTGGGCCGCGCCACCGGCTCGGAAGGCGTCATGGGC GGCCAGATGGTGGACATCGCTTCAGAGGGCGACCCG TCGATCGACCTCCAGACGCTCGAGTGGATTCACATCC ACAAGACCGCGATGCTCTTGGAATGCTCCGTCGTCTG CGGAGCAATCATCGGCGGCGCGTCGGAAATCGTTATC GAGCGCGCTCGCCGCTACGCCCGCTGCGTCGGCCTC CTCTTCCAGGTCGTGGACGACATCCTCGACGTCACAA AGTCGTCGGACGAGTTGGGTAAAACGGCGGGCAAGG ATCTCATTTCGGACAAGGCGACCTACCCTAAGCTCATG GGCTTGGAGAAGGCCAAGGAATTCAGCGACGAGCTTC TCAACCGCGCCAAGGGAGAGTTGTCGTGCTTCGACCC AGTCAAGGCCGCGCCGCTCCTCGGCCTCGCTGATTAC GTCGCCTTCCGCCAAAACTAGTCTACGAGACCC	*Abies grandis*	Kirby et al., 2021
NDPS1	GGGTCTCGAATGTCGGCCCGAGGCCTTAACAAAATCT CGTGCTCCCTCAACCTCCAGACTGAGAAGCTCTGCTA CGAGGACAACGACAACGACCTCGACGAGGAGCTCATG CCGAAGCACATCGCGCTCATCATGGACGGCAACCGCC GCTGGGCCAAGGACAAGGGATTGGAGGTCTACGAGG GCCATAAGCACATCATTCCCAAGTTGAAGGAGATCTGC GACATCTCGTCGAAGCTCGGGATCCAGATCATCACGG CTTTCGCGTTCAGCACCGAGAACTGGAAGCGCTCGAA GGAAGAGGTCGACTTCCTCCTCCAGATGTTTGAGGAG ATCTACGACGAGTTCTCTCGGAGCGGCGTCCGTGTGT CCATCATCGGCTGCAAGTCAGACCTCCCCATGACCCT	*Solanum lycopersicum*	Wei et al., 2020

	CCAGAAGTGCATCGCCCTCACCGAGGAAACCACAAAG GGCAACAAGGGTCTTCACCTCGTCATTGCCCTGAACTA CGGCGGATACTACGACATTCTCCAAGCAACCAAGTCG ATCGTCAACAAGGCGATGAACGGACTTCTCGACGTTG AGGACATCAACAAGAACCTGTTCGACCAAGAATTGGAA TCCAAGTGCCCAAACCCTGACCTCCTCATCCGCACCG GCGGCGAGCAGCGCGTCTCGAACTTCCTCTTGTGGCA GCTTGCGTACACGGAGTTCTATTTTACGAACACGCTCT TCCCGGACTTCGGTGAAGAGGATCTCAAGGAGGCTAT CATGAATTTCCAGCAGCGCCACAGGCGCTTCGGCGGC CACACGTACTAGTCTACGAGACCC		
tHMG1	ACGGGGTCTCCAATGACCCAGTCGGTCAAGGTGGTCG AGAAGCACGTCCCGATCGTCATCGAGAAGCCAAGCGA GAAGGAGGAGGACACCTCGTCGGAAGACTCCATCGAG CTCACGGTCGGAAAACAGCCCAAGCCCGTCACCGAGA CACGCTCTTTGGACGACCTCGAGGCTATCATGAAGGC AGGCAAGACCAAGCTCCTGGAGGACCACGAGGTCGTC AAGCTCAGCCTCGAGGGCAAGCTCCCTTTGTATGCGC TTGAGAAGCAGCTCGGGGACAACACCCGCGCTGTCG GCATCCGCCGCTCGATCATCTCGCAGCAGTCTAATAC CAAGACATTGGAAACGTCGAAGCTTCCGTACCTCCACT ACGACTACGACCGCGTTTTCGGAGCCTGTTGCGAGAA CGTCATTGGCTACATGCCCCTCCCCGTCGGCGTTGCG GGCCCCATGAACATCGACGGCAAGAACTACCACATTC CTATGGCCACCACGGAGGGCTGCCTTGTCGCCTCAAC CATGCGCGGCTGCAAGGCCATCAACGCCGGCGGCGG GGTCACCACGGTCCTCACTCAGGACGGCATGACGCGA GGTCCGTGCGTATCCTTCCCCTCGCTCAAGCGCGCGG GAGCCGCGAAGATCTGGCTCGACTCCGAGGAGGGCT TGAAGTCCATGCGCAAGGCCTTCAACTCGACGTCGCG CTTTGCGCGTCTCCAGTCGCTCCACTCGACCCTCGCG GGCAACCTCCTCTTCATCAGGTTCCGCACCACCACGG GCGACGCCATGGGCATGAACATGATCTCCAAGGGCGT CGAACACTCGCTCGCCGTCATGGTCAAGGAGTACGGC TTCCCTGACATGGACATTGTCAGCGTCTCGGGAAACTA CTGCACGGACAAGAAGCCAGCAGCGATCAACTGGATC GAAGGCCGAGGCAAGAGTGTCGTGGCCGAAGCCACC ATCCCCGCGCACATCGTCAAGTCGGTTCTCAAAAGCG AAGTCGACGCTCTCGTCGAGCTCAACATCAGCAAGAA CCTGATCGGTTCGGCCATGGCTGGCTCGGTGGGAGG CTTCAACGCACACGCGGCAAACCTCGTCACCGCCATC TACCTCGCGACTGGCCAGGATCCGGCTCAGAATGTCG AGTCGTCCAACTGCATCACGCTCATGAGCAACGTCGA CGGGAACCTGCTCATCTCCGTCTCGATGCCGTCGATC GAGGTCGGCACCATTGGTGGAGGTACTATCTTGGAGC	*Yarrowia lipolytica*	

	CACAGGGGGCGATGCTGGAGATGCTCGGCGTCCGCG GCCCACACATCGAGACGCCGGGTGCCAACGCCCAAC AGCTTGCGCGCATCATCGCTTCGGGAGTTCTCGCAGC GGAGCTTTCGCTCTGCTCGGCGCTCGCTGCGGGCCAT CTCGTGCAAAGCCATATGACGCACAACCGGTCCCAGG CGCCTACCCCGGCCAAGCAGTCTCAGGCCGATCTCCA GAGACTCCAAAACGGCTCGAACATTTGCATCCGGTCAT AGTCTACGAGACCCCGT		
HYP3	GGGTCTCGCGTCCTAGATGCCCCGGAGTCCCGGCTTG GCGTAGCGCAAGCAGGTGGCTGAGTAAAAGATGTTAC CGGCGCAGACCTCGACCATTCCGCGAGCAAAGCGCC ATTGCGAAGGCGTGAGGTCGGTGCCCTTGAGGCGCG TCAGCTCGTCGTGGATCTGCAATTCGAGGTCCCAGAG AAATGCGCGGAGGAGGACCTTAGCCGAGCGCGTAGA GGTATCCACGAGCTGCTCCAAGACGCGGACCGCGTTG ATGACGGCGCTGCCAGTCTCCTGCTCGGCGACGACTT CCTTGTTGAACGAGTAGAGGTCGTTGGTGAGCGAGAA ATGCGTCATGTAGAGCCGGTGAAGCTTGCGGAGGGG CTCCATGTCCGCTGGGGTGAGGAAGATCTCGCACGTG AGCAAGATCGCGGCGTCGCAGAAGTCCACGGCGAAG TCGCCGATGCGGTAGTCAATGTACTCGTTGAACGATCT GAAGGTCATCGAAGCGGGGAGGGCCTTGCGAGTGTG CTCAACCCAATTGAGGATGCCTTGGAGAAACTGCGGC CCGACGACCGGGTCCTCTCGGTTTGTCTCCTCCGAGA ACTCCTTGAAAATGCGGTTTTGCCAGAGAATGTCGGC GCCCTTTCCGTTAGGGGTGGAGGTGTCCGCGAGGGC GCTGTCGATGATCGTGGACTGGACCGACTGGTACTCG CAGACGTCGTCGTGCATGAAGATGAGGAGGTTCGCCT TGGCCATGATGCGCGCACGCGGAGCGTTAGCTGCAG GGAACATGTTCATGGGGGCCGAGACGGAGGTGTCGA GAAGCTCTCGGCGCTTGGCCGCCTTCTTATCAGAGAC GGCCAGTTCTGCCGGAAGCGAACCCTGCTCGCCAGG CGCCGCCGCGACGATCTGGTCCATCAGCTCCGCCGCT GCTTCCTCGGCGTCCTTCCAGTACTTCGACTGCGCCG CAGCTGGGAACGACGTTGGCCACGGAACGCCGAGCG ACCTGGGGTCGAATACGTTGCAGTTCTCAAGGCCGGG GACGGAGCGGGAGATAGCTTCCCGGAGGAACTCGAA GTTCTCCTGTTTGCTCTCGGTCTGGAAGCTGATGCCGA CCGGGTCGAACGAGCACGTGATGGGCCTCATCGCAC GAGACCC		Kirby et al., 2021

IV. GENERAL CONCLUSIONS

The results of the first chapter of this thesis demonstrated a bioinformatics workflow developed to detect novel *cis*-regulatory elements from *R. toruloides* transcriptomic data, when grown simulating industry conditions. In the second chapter, the framework of synthetic biology combined with metabolic engineering technologies were applied to *R. toruloides* for engineering the endogenous mevalonate pathway to produce non-native monoterpenes. Although none of the monoterpenes were detected, the DBTL cycle was successfully applied to understand what can be improved for future endeavors.

Altogether, the two chapters created new perspectives for engineering *R. toruloides* for fine chemical production. As depicted in the first chapter, the terpenoid backbone biosynthesis pathway was shown to be up-regulated in the condition using sugarcane as substrate, which indicates that this condition can be used for further engineering and production of non-native terpenoids, such as monoterpenes. Studies have already shown that combinatorial expression of key pathway genes together with transcriptional activators can improve terpene production. Thus, the transcription factors found with the motif discovery tool can also be combined with metabolic engineering strategies. For instance, HAC1 and ADR1 were predicted as positive transcriptional regulators responding to cell stress and ethanol, and adding their motifs in the promoter sequence of mevalonate pathway genes would possibly enhance enzyme expression at the transcriptional level. Similarly, RPN4 was predicted as a negative transcriptional regulator, and adding its motif to promoter sequencing of enzymes competing for the substrates of interest can also be an option. Moreover, the data generated here can also

be implemented into metabolic models in the future, helping to simplify gene editing through smarter decisions.

The results of this thesis have provided resourceful information regarding the fusion of synthetic biology, bioinformatics and metabolic engineering applied to *R. toruloides,* as the urgent need for such a robust microorganism in biorefineries is becoming clearer. This work is a small but meaningful contribution to confirming the role of this oleaginous, carotenogenic yeast as a sustainable microbial cell factory in the circular bioeconomy.

V. REFERENCES

Abdel-Banat, B. M. A. *et al.* (2010) ‘High-temperature fermentation: How can processes for ethanol production at high temperatures become superior to the traditional process using mesophilic yeast?’, *Applied Microbiology and Biotechnology*, 85(4), pp. 861–867. doi: 10.1007/s00253-009-2248-5.

Abomohra, A. E. F. *et al.* (2016) ‘Microalgal biomass production as a sustainable feedstock for biodiesel: Current status and perspectives’, *Renewable and Sustainable Energy Reviews*, 64, pp. 596–606. doi: 10.1016/j.rser.2016.06.056.

Amarelle, V. *et al.* (2018) ‘Expanding the toolbox of broad host-range transcriptional terminators for Proteobacteria through metagenomics’, *bioRxiv*. doi: 10.1101/485938.

Andrews, S. (2018) *FastQC: A Quality Control tool for High Throughput Sequence Data*, *FastQC: A Quality Control tool for High Throughput Sequence Data*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Antonieto, A. C. C. *et al.* (2019) ‘A Novel Cys2His2 Zinc Finger Homolog of AZF1 Modulates Holocellulase Expression in Trichoderma reesei’, *mSystems*. Edited by C. Vickers, 4(4), pp. 1–14. doi: 10.1128/mSystems.00161-19.

Appleton, E. *et al.* (2017) ‘Design automation in synthetic biology’, *Cold Spring Harbor Perspectives in Biology*, 9(4). doi: 10.1101/cshperspect.a023978.

Bastian, M., Heymann, S. and Jacomy, M. (2009) ‘Gephi: An Open Source Software for Exploring and Manipulating Networks’, *Proceedings of the Third International ICWSM Conference*, pp. 361–362. Available at: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Bastos, V. D. (2007) ‘Etanol, alcoolquímica e biorrefinarias’, *BNDES Setorial*, 1(25), pp. 5–38. doi: 10.1017/CBO9781107415324.004.

Benjamin, K. R. *et al.* (2016) ‘Developing commercial production of semi-synthetic artemisinin, and of β-Farnesene, an Isoprenoid Produced by Fermentation of Brazilian Sugar’, *Journal of the Brazilian Chemical Society*, 27(8), pp. 1339–1345. doi: 10.5935/0103-5053.20160119.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) ‘Trimmomatic: A flexible trimmer for Illumina sequence data’, *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Bonturi, N. *et al.* (2017) ‘Microbial oil production in sugarcane bagasse hemicellulosic

hydrolysate without nutrient supplementation by a Rhodosporidium toruloides adapted strain', *Process Biochemistry*, 57, pp. 16–25. doi: 10.1016/j.procbio.2017.03.007.
Bonturi, N., Pinheiro, M. J., Monteiro de Oliveira, P., *et al.* (2022) 'Development of a dedicated Golden Gate Assembly Platform (RtGGA) for Rhodotorula toruloides', *Metabolic Engineering Communications*, 15(May), p. e00200. doi: 10.1016/j.mec.2022.e00200.
Bonturi, N., Pinheiro, M. J., de Oliveira, P. M., *et al.* (2022) 'Development of a dedicated Golden Gate Assembly Platform (RtGGA) for Rhodotorula toruloides', *Metabolic Engineering Communications*, 15(March). doi: 10.1016/j.mec.2022.e00200.
Borodina, I. and Nielsen, J. (2014) 'Advances in metabolic engineering of yeast Saccharomyces cerevisiae for production of chemicals', *Biotechnology Journal*, 9(5), pp. 609–620. doi: 10.1002/biot.201300445.
Bray, N. L. *et al.* (2016) 'Near-optimal probabilistic RNA-seq quantification', *Nature Biotechnology*, 34(5), pp. 525–527. doi: 10.1038/nbt.3519.
Bredeweg, E. L. *et al.* (2017) 'A molecular genetic toolbox for Yarrowia lipolytica', *Biotechnology for Biofuels*, 10(1), pp. 1–22. doi: 10.1186/s13068-016-0687-7.
Bubis, J. A. *et al.* (2020) 'Rpn4 and proteasome-mediated yeast resistance to ethanol includes regulation of autophagy', *Applied Microbiology and Biotechnology*, 104(9), pp. 4027–4041. doi: 10.1007/s00253-020-10518-x.
Campbell, K., Xia, J. and Nielsen, J. (2017) 'The Impact of Systems Biology on Bioprocessing', pp. 1156–1168.
Cao, M. *et al.* (2022) 'Metabolic engineering of oleaginous yeast Rhodotorula toruloides for overproduction of triacetic acid lactone', *Biotechnology and Bioengineering*, 119(9), pp. 2529–2540. doi: 10.1002/bit.28159.
Celińska, E. *et al.* (2017) 'Golden Gate Assembly system dedicated to complex pathway manipulation in Yarrowia lipolytica', *Microbial Biotechnology*, 10(2), pp. 450–455. doi: 10.1111/1751-7915.12605.
Cherubini, F. (2010) 'The biorefinery concept: Using biomass instead of oil for producing energy and chemicals', *Energy Conversion and Management*, 51(7), pp. 1412–1421. doi: 10.1016/j.enconman.2010.01.015.
Cho, J. S. *et al.* (2022) 'Designing Microbial Cell Factories for the Production of

Chemicals', *Journal of the American Chemical Society*, 2(8), pp. 1781–1799. doi: 10.1021/jacsau.2c00344.

Cohen, S. N. *et al.* (1973) 'Construction of Biologically Functional Bacterial Plasmids In Vitro', *Proceedings of the National Academy of Sciences*, 70(11), pp. 3240–3244. doi: 10.1073/pnas.70.11.3240.

Colussi, P. A. and Taron, C. H. (2005) 'Kluyveromyces lactis LAC4 promoter variants that lack function in bacteria but retain full function in K. lactis', *Applied and Environmental Microbiology*, 71(11), pp. 7092–7098. doi: 10.1128/AEM.71.11.7092-7098.2005.

Coradetti, S. T. *et al.* (2018) 'Functional genomics of lipid metabolism in the oleaginous yeast Rhodosporidium toruloides', *eLife*, 7, p. e32110. doi: 10.7554/eLife.32110.

Curran, K. A. *et al.* (2013) 'Use of expression-enhancing terminators in Saccharomyces cerevisiae to increase mRNA half-life and improve gene expression control for metabolic engineering applications', *Metabolic Engineering*, 19, pp. 88–97. doi: 10.1016/j.ymben.2013.07.001.

Darvishi, F. *et al.* (2018) 'Advances in synthetic biology of oleaginous yeast Yarrowia lipolytica for producing non-native chemicals', *Applied Microbiology and Biotechnology*, 102(14), pp. 5925–5938. doi: 10.1007/s00253-018-9099-x.

Denby, C. M. *et al.* (2018) 'Industrial brewing yeast engineered for the production of primary flavor determinants in hopped beer', *Nature Communications*, 9(1), pp. 1–10. doi: 10.1038/s41467-018-03293-x.

Devi, A. *et al.* (2022) 'Lignocellulosic Biomass Valorization for Bioethanol Production: a Circular Bioeconomy Approach', *Bioenergy Research*. doi: 10.1007/s12155-022-10401-9.

Dinh, H. V. *et al.* (2019) 'A comprehensive genome-scale model for Rhodosporidium toruloides IFO0880 accounting for functional genomics and phenotypic data', *Metabolic Engineering Communications*, 9(August), p. e00101. doi: 10.1016/j.mec.2019.e00101.

Dutta, N. *et al.* (2022) 'An Insight into Valorization of Lignocellulosic Biomass by Optimization with the Combination of Hydrothermal (HT) and Biological Techniques: A Review', *Sustainable Chemistry*, 3(1), pp. 35–55. doi: 10.3390/suschem3010003.

Feng, Y. *et al.* (2020) 'Crystal Structure of Geranylgeranyl Pyrophosphate Synthase

(CrtE) Involved in Cyanobacterial Terpenoid Biosynthesis', *Frontiers in Plant Science*, 11(May), pp. 1–14. doi: 10.3389/fpls.2020.00589.

Fernandes, E. S. *et al.* (2007) 'Anti-inflammatory effects of compounds alpha-humulene and (−)-trans-caryophyllene isolated from the essential oil of Cordia verbenacea', *European Journal of Pharmacology*, 569(3), pp. 228–236. doi: 10.1016/j.ejphar.2007.04.059.

Fletcher, E. *et al.* (2017) 'Evolutionary engineering reveals divergent paths when yeast is adapted to different acidic environments', *Metabolic Engineering*, 39(September 2016), pp. 19–28. doi: 10.1016/j.ymben.2016.10.010.

Gauthier, J. *et al.* (2019) 'A brief history of bioinformatics', *Briefings in Bioinformatics*, 20(6), pp. 1981–1996. doi: 10.1093/bib/bby063.

Geiselman, G. M. *et al.* (2020) 'Production of ent-kaurene from lignocellulosic hydrolysate in Rhodosporidium toruloides', *Microbial Cell Factories*, 19(1), pp. 1–12. doi: 10.1186/s12934-020-1293-8.

Gibson, D. G. *et al.* (2009) 'Enzymatic assembly of DNA molecules up to several hundred kilobases.', *Nature methods*, 6(5), pp. 343–5. doi: 10.1038/nmeth.1318.

Gong, Z., Nielsen, J. and Zhou, Y. J. (2017) 'Engineering Robustness of Microbial Cell Factories', *Biotechnology Journal*, 12(10). doi: 10.1002/biot.201700014.

Guo, Z. and Sherman, F. (1996) '3'-end-forming signals of yeast mRNA', *Trends in Biochemical Sciences*, 21(12), pp. 477–481. doi: 10.1016/S0968-0004(96)10057-8.

Gupta, S. *et al.* (2007) 'Quantifying similarity between motifs', *Genome Biology*, 8(2). doi: 10.1186/gb-2007-8-2-r24.

Heinz, S. *et al.* (2010) 'Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities', *Molecular Cell*, 38(4), pp. 576–589. doi: 10.1016/j.molcel.2010.05.004.

Huang, C. J. *et al.* (2018) 'Experimental Evolution of Yeast for High-Temperature Tolerance', *Molecular biology and evolution*, 35(8), pp. 1823–1839. doi: 10.1093/molbev/msy077.

Hwang, H. J., Lee, S. Y. and Lee, P. C. (2018) 'Engineering and application of synthetic nar promoter for fine-tuning the expression of metabolic pathway genes in Escherichia coli', *Biotechnology for Biofuels*, 11(1), pp. 1–13. doi: 10.1186/s13068-018-1104-1.

Inukai, S., Kock, K. H. and Bulyk, M. L. (2017) ‘Transcription factor–DNA binding: beyond binding site motifs’, *Current Opinion in Genetics & Development*, 43(3), pp. 110–119. doi: 10.1016/j.gde.2017.02.007.

Ito, Y. *et al.* (2016) ‘Enhancement of protein production via the strong DIT1 terminator and two RNA-binding proteins in Saccharomyces cerevisiae’, *Scientific Reports*, 6(October), pp. 1–9. doi: 10.1038/srep36997.

Jamar, N. H., Kritsiligkou, P. and Grant, C. M. (2017) ‘The non-stop decay mRNA surveillance pathway is required for oxidative stress tolerance’, *Nucleic Acids Research*, 45(11), pp. 6881–6893. doi: 10.1093/nar/gkx306.

Jiang, Z., Kempinski, C. and Chappell, J. (2016) ‘Extraction and Analysis of Terpenes/ Terpenoids Public Access’, *Current Protocols in Plant Biology*, 25(3), pp. 289–313. doi: 10.1002/cppb.20024.Extraction.

Jiao, X. *et al.* (2019) ‘Developing a CRISPR/Cas9 System for Genome Editing in the Basidiomycetous Yeast Rhodosporidium toruloides’, *Biotechnology Journal*, 14(7), p. 1900036. doi: 10.1002/biot.201900036.

Johns, Alexander M.B., Love, J. and Aves, S. J. (2016) ‘Four inducible promoters for controlled gene expression in the oleaginous yeast Rhodotorula toruloides’, *Frontiers in Microbiology*, 7(OCT), pp. 1–12. doi: 10.3389/fmicb.2016.01666.

Johns, Alexander M B, Love, J. and Aves, S. J. (2016) ‘Four inducible promoters for controlled gene expression in the oleaginous yeast Rhodotorula toruloides’, *Frontiers in Microbiology*, 7(OCT), pp. 1–12. doi: 10.3389/fmicb.2016.01666.

Kang, A. and Lee, T. S. (2015) ‘Converting Sugars to Biofuels: Ethanol and Beyond’, *Bioengineering*, 2(4), pp. 184–203. doi: 10.3390/bioengineering2040184.

Kim, J. *et al.* (2021) ‘Multi-Omics Driven Metabolic Network Reconstruction and Analysis of Lignocellulosic Carbon Utilization in Rhodosporidium toruloides’, *Frontiers in Bioengineering and Biotechnology*, 8(January). doi: 10.3389/fbioe.2020.612832.

Kirby, J. *et al.* (2021) ‘Further engineering of R . toruloides for the production of terpenes from lignocellulosic biomass’, *Biotechnology for Biofuels*, pp. 1–16. doi: 10.1186/s13068-021-01950-w.

Konur, S. *et al.* (2021) ‘Toward Full-Stack in Silico Synthetic Biology: Integrating Model Specification, Simulation, Verification, and Biological Compilation’, *ACS Synthetic*

Biology, 10(8), pp. 1931–1945. doi: 10.1021/acssynbio.1c00143.
Küberl, A. *et al.* (2011) 'High-quality genome sequence of Pichia pastoris CBS7435', *Journal of Biotechnology*, 154(4), pp. 312–320. doi: 10.1016/j.jbiotec.2011.04.014.
Lambertz, C. *et al.* (2016) 'Kluyveromyces lactis as an Expression Host for Enzymes that Degrade Lignocellulosic Biomass', *Cellulose Chem. Technol*, 50, pp. 3–4.
Larroude, M. *et al.* (2019) 'A modular Golden Gate toolkit for Yarrowia lipolytica synthetic biology', *Microbial Biotechnology*, 12(6), pp. 1249–1259. doi: 10.1111/1751-7915.13427.
Legault, J. and Pichette, A. (2007) 'Potentiating effect of β-caryophyllene on anticancer activity of α-humulene, isocaryophyllene and paclitaxel', *Journal of Pharmacy and Pharmacology*, 59(12), pp. 1643–1647. doi: 10.1211/jpp.59.12.0005.
Li, M. *et al.* (2021) 'CRISPR-mediated multigene integration enables Shikimate pathway refactoring for enhanced 2-phenylethanol biosynthesis in Kluyveromyces marxianus', *Biotechnology for Biofuels*, 14(1), pp. 1–15. doi: 10.1186/s13068-020-01852-3.
Li, M. Z. and Elledge, S. J. (2007) 'Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC.', *Nature methods*, 4(3), pp. 251–6. doi: 10.1038/nmeth1010.
Li, R. *et al.* (2021) 'Production of plant volatile terpenoids (rose oil) by yeast cell factories', *Green Chemistry*, 23(14), pp. 5088–5096. doi: 10.1039/d1gc00917f.
Li, W. *et al.* (2022) 'Advances in Metabolic Engineering Paving the Way for the Efficient Biosynthesis of Terpenes in Yeasts', *Journal of Agricultural and Food Chemistry*, 70(30), pp. 9246–9261. doi: 10.1021/acs.jafc.2c03917.
Lin, X. *et al.* (2014) 'Functional integration of multiple genes into the genome of the oleaginous yeast Rhodosporidium toruloides', *FEMS Yeast Research*, 14(4), pp. 547–555. doi: 10.1111/1567-1364.12140.
Liobikiene, G. *et al.* (2020) 'The trends in bioeconomy development in the European Union: Exploiting capacity and productivity measures based on the land footprint approach', *Land Use Policy*, 91(November 2019), p. 104375. doi: 10.1016/j.landusepol.2019.104375.
Litwin, I. *et al.* (2013) 'Oxidative Stress and Replication-Independent DNA Breakage Induced by Arsenic in Saccharomyces cerevisiae', *PLoS Genetics*, 9(7). doi:

10.1371/journal.pgen.1003640.
Liu, H. *et al.* (2017) ‘Fast and efficient genetic transformation of oleaginous yeast Rhodosporidium toruloides by using electroporation’, *FEMS Yeast Research*, 17(2), pp. 1–11. doi: 10.1093/femsyr/fox017.
Liu, S. *et al.* (2021) ‘Engineering Rhodosporidium toruloides for limonene production’, *Biotechnology for Biofuels*, 14(1), pp. 1–11. doi: 10.1186/s13068-021-02094-7.
Liu, X. *et al.* (2019) ‘RNA interference in the oleaginous yeast Rhodosporidium toruloides’, *FEMS Yeast Research*, 19(3), pp. 1–10. doi: 10.1093/femsyr/foz031.
Liu, Y. *et al.* (2013) ‘Characterization of glyceraldehyde-3-phosphate dehydrogenase gene RtGPD1 and development of genetic transformation method by dominant selection in oleaginous yeast Rhodosporidium toruloides’, *Applied Microbiology and Biotechnology*, 97(2), pp. 719–729. doi: 10.1007/s00253-012-4223-9.
Liu, Y. *et al.* (2016a) ‘Developing a set of strong intronic promoters for robust metabolic engineering in oleaginous Rhodotorula (Rhodosporidium) yeast species’, *Microbial Cell Factories*, 15(1), pp. 1–9. doi: 10.1186/s12934-016-0600-x.
Liu, Y. *et al.* (2016b) ‘Developing a set of strong intronic promoters for robust metabolic engineering in oleaginous Rhodotorula (Rhodosporidium) yeast species’, *Microbial Cell Factories*, 15(1), pp. 1–9. doi: 10.1186/s12934-016-0600-x.
Liu, Y. *et al.* (2021) ‘Biofuels for a sustainable future’, *Cell*, 184(6), pp. 1636–1647. doi: 10.1016/j.cell.2021.01.052.
Löbs, A. K. *et al.* (2017) ‘CRISPR-Cas9-enabled genetic disruptions for understanding ethanol and ethyl acetate biosynthesis in Kluyveromyces marxianus’, *Biotechnology for Biofuels*, 10(1), pp. 1–14. doi: 10.1186/s13068-017-0854-5.
Lõoke, M., Kristjuhan, K. and Kristjuhan, A. (2017) ‘Extraction of genomic DNA from yeasts for PCR-based applications’, *BioTechniques*, 62(1), p. v. doi: 10.2144/000114497.
Lopes, H. J. S. (2020) ‘Estratégias de adaptação e de cultivo de Rhodotorula torloides visando produção de óleo microbiano a partir de hidrolisados hemicelulósicos’, *Universidade Estadual de Campinas*, 21(1), pp. 1–9. Available at: http://mpoc.org.my/malaysian-palm-oil-industry/.
de Lorenzo, V. (2011) ‘Beware of metaphors chasses and orthogonality in synthetic

biology', *Bioengineered Bugs*, 2(1), pp. 3–7. doi: 10.4161/bbug.2.1.13388.
Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), pp. 1–21. doi: 10.1186/s13059-014-0550-8.
Lucero, P. *et al.* (2000) 'Internal trehalose protects endocytosis from inhibition by ethanol in Saccharomyces cerevisiae', *Applied and Environmental Microbiology*, 66(10), pp. 4456–4461. doi: 10.1128/AEM.66.10.4456-4461.2000.
Luo, W. *et al.* (2009) 'GAGE: Generally applicable gene set enrichment for pathway analysis', *BMC Bioinformatics*, 10, pp. 1–17. doi: 10.1186/1471-2105-10-161.
Luo, W. and Brouwer, C. (2013) 'Pathview: An R/Bioconductor package for pathway-based data integration and visualization', *Bioinformatics*, 29(14), pp. 1830–1831. doi: 10.1093/bioinformatics/btt285.
Martins-Santana, L. *et al.* (2018) 'Systems and Synthetic Biology Approaches to Engineer Fungi for Fine Chemical Production', *Frontiers in Bioengineering and Biotechnology*, 6(October), pp. 1–18. doi: 10.3389/fbioe.2018.00117.
Martins, L. C. *et al.* (2021) 'Complete utilization of the major carbon sources present in sugar beet pulp hydrolysates by the oleaginous red yeasts rhodotorula toruloides and r. Mucilaginosa', *Journal of Fungi*, 7(3). doi: 10.3390/jof7030215.
Meng, L. *et al.* (2021) 'Effect of overexpression of SNF1 on the transcriptional and metabolic landscape of baker's yeast under freezing stress', *Microbial Cell Factories*, 20(1), pp. 1–13. doi: 10.1186/s12934-020-01503-0.
Monteiro de Oliveira, P. *et al.* (2021) 'Screening and Growth Characterization of Non-conventional Yeasts in a Hemicellulosic Hydrolysate', *Frontiers in Bioengineering and Biotechnology*, 9(April), pp. 1–13. doi: 10.3389/fbioe.2021.659472.
Monteiro, J. L. F. and Veloso, C. O. (2004) 'Catalytic conversion of terpenes into fine chemicals', *Topics in Catalysis*, 27(1–4), pp. 169–180. doi: 10.1023/B:TOCA.0000013551.99872.8d.
Monteiro, L. M. O. *et al.* (2020) 'Unraveling the Complex Interplay of Fis and IHF Through Synthetic Promoter Engineering', *Frontiers in Bioengineering and Biotechnology*, 8(June), pp. 1–13. doi: 10.3389/fbioe.2020.00510.
Näätsaari, L. *et al.* (2012) 'Deletion of the pichia pastoris ku70 homologue facilitates

platform strain generation for gene expression and synthetic biology', *PLoS ONE*, 7(6). doi: 10.1371/journal.pone.0039720.

Nambu-Nishida, Y. *et al.* (2017) 'Development of a comprehensive set of tools for genome engineering in a cold- A nd thermo-tolerant Kluyveromyces marxianus yeast strain', *Scientific Reports*, 7(1), pp. 1–7. doi: 10.1038/s41598-017-08356-5.

Nielsen, J. and Keasling, J. D. (2016) 'Engineering Cellular Metabolism', *Cell*, 164(6), pp. 1185–1197. doi: 10.1016/j.cell.2016.02.004.

Niu, F. X. *et al.* (2018) 'Enhancing Production of Pinene in Escherichia coli by Using a Combination of Tolerance, Evolution, and Modular Co-culture Engineering', *Frontiers in Microbiology*, 9. doi: 10.3389/fmicb.2018.01623.

Nora, L. C. *et al.* (2018) 'The art of vector engineering: towards the construction of next-generation genetic tools', *Microbial Biotechnology*, 0(0), pp. 1–23. doi: 10.1111/1751-7915.13318.

Nora, L. C., Wehrs, M., Kim, J., Skerker, J., *et al.* (2019) 'A toolset of constitutive promoters for engineering Rhodosporidium toruloides', *Microbial Cell Factories*, p. 2.

Nora, L. C., Wehrs, M., Kim, J., Cheng, J. F., *et al.* (2019) 'A toolset of constitutive promoters for metabolic engineering of Rhodosporidium toruloides', *Microbial Cell Factories*, 18(1), pp. 1–11. doi: 10.1186/s12934-019-1167-0.

Nora, L. C., Westmann, C. A., *et al.* (2019) 'Recent advances in plasmid-based tools for establishing novel microbial chassis', *Biotechnology Advances*, 37(8), p. 107433. doi: 10.1016/j.biotechadv.2019.107433.

Obst, U., Lu, T. K. and Sieber, V. (2017) 'A Modular Toolkit for Generating Pichia pastoris Secretion Libraries', *ACS Synthetic Biology*, 6(6), pp. 1016–1025. doi: 10.1021/acssynbio.6b00337.

Van Ooyen, A. J. J. *et al.* (2006) 'Heterologous protein production in the yeast Kluyveromyces lactis', *FEMS Yeast Research*, 6(3), pp. 381–392. doi: 10.1111/j.1567-1364.2006.00049.x.

Opgenorth, P. *et al.* (2019) 'Lessons from Two Design-Build-Test-Learn Cycles of Dodecanol Production in Escherichia coli Aided by Machine Learning', *ACS Synthetic Biology*, 8(6), pp. 1337–1351. doi: 10.1021/acssynbio.9b00020.

Oswald, M. *et al.* (2007) 'Monoterpenoid biosynthesis in Saccharomyces cerevisiae',

FEMS Yeast Research, 7(3), pp. 413–421. doi: 10.1111/j.1567-1364.2006.00172.x.
Otoupal, P. B. *et al.* (2019) 'Multiplexed CRISPR-Cas9 based genome editing of Rhodosporidium toruloides'.
Pacheco, T. F. (2011) 'Produção de Etanol: Primeira ou Segunda Geração?', *2011*, 1, p. 6.
Paduch, R. *et al.* (2007) 'Terpenes: Substances useful in human healthcare', *Archivum Immunologiae et Therapiae Experimentalis*, 55(5), pp. 315–327. doi: 10.1007/s00005-007-0039-1.
Palahí, M. and Adams, J. (2020) *Why The World Needs A 'circular Bioeconomy' - For Jobs, Biodiversity And Prosperity*. Available at: https://www.forbes.com/sites/worldeconomicforum/2020/10/06/why-the-world-needs-a-circular-bioeconomyfor-jobs-biodiversity-and-prosperity/?sh=601148c053a8.
Pan, R. *et al.* (2011) 'Sequential deletion of Pichia pastoris genes by a self-excisable cassette', *FEMS Yeast Research*, 11(3), pp. 292–298. doi: 10.1111/j.1567-1364.2011.00716.x.
Pardo, E. *et al.* (2015) 'De novo production of six key grape aroma monoterpenes by a geraniol synthase-engineered S. cerevisiae wine strain', *Microbial Cell Factories*, 14(1), pp. 1–8. doi: 10.1186/s12934-015-0306-5.
Park, Y. K., Nicaud, J. M. and Ledesma-Amaro, R. (2017) 'The Engineering Potential of Rhodosporidium toruloides as a Workhorse for Biotechnological Applications', *Trends in Biotechnology*, 36(3), pp. 304–317. doi: 10.1016/j.tibtech.2017.10.013.
Park, Y. K., Nicaud, J. M. and Ledesma-Amaro, R. (2018) 'The Engineering Potential of Rhodosporidium toruloides as a Workhorse for Biotechnological Applications', *Trends in Biotechnology*, 36(3), pp. 304–317. doi: 10.1016/j.tibtech.2017.10.013.
Peplow, M. (2016) 'Synthetic biology's first malaria drug meets market resistance', *Nature*, 530(7591), pp. 389–390. doi: 10.1038/530390a.
Pichette, A. *et al.* (2006) 'Composition and antibacterial activity of Abies balsamea essential oil', *Phytotherapy Research*, 20(5), pp. 371–373. doi: 10.1002/ptr.1863.
Pinheiro, M. J. *et al.* (2020) 'Xylose Metabolism and the Effect of Oxidative Stress on Lipid and Carotenoid Production in Rhodotorula toruloides: Insights for Future Biorefinery', *Frontiers in Bioengineering and Biotechnology*, 8(August), pp. 1–15. doi:

10.3389/fbioe.2020.01008.
de Pourcq, K. *et al.* (2012) ‘Engineering Yarrowia lipolytica to produce glycoproteins homogeneously modified with the universal man3GlcNAc2 N-glycan core’, *PLoS ONE*, 7(6). doi: 10.1371/journal.pone.0039976.
Prielhofer, R. *et al.* (2017) ‘GoldenPiCS: A Golden Gate-derived modular cloning system for applied synthetic biology in the yeast Pichia pastoris’, *BMC Systems Biology*, 11(1), pp. 1–14. doi: 10.1186/s12918-017-0492-3.
Quan, J. and Tian, J. (2009) ‘Circular polymerase extension cloning of complex gene libraries and pathways.’, *PloS one*, 4(7), p. e6441. doi: 10.1371/journal.pone.0006441.
R Core Team (2019) *A Language and Environment for Statistical Computing*, *R Foundation for Statistical Computing*. Available at: http://www.r-project.org.
Rajkumar, A. S. and Morrissey, J. P. (2020) ‘Rational engineering of Kluyveromyces marxianus to create a chassis for the production of aromatic products’, *Microbial Cell Factories*, 19(1), pp. 1–19. doi: 10.1186/s12934-020-01461-7.
Rajkumar, A. S. and Morrissey, J. P. (2022) ‘Protocols for marker-free gene knock-out and knock-down in Kluyveromyces marxianus using CRISPR/Cas9’, *FEMS Yeast Research*, 22(1), pp. 1–16. doi: 10.1093/femsyr/foab067.
Reece-Hoyes, J. S. and Walhout, A. J. M. (2018) ‘Gateway recombinational cloning’, *Cold Spring Harbor Protocols*, 2018(1), pp. 1–6. doi: 10.1101/pdb.top094912.
Ribeiro, O. *et al.* (2007) ‘Application of the Cre-loxP system for multiple gene disruption in the yeast Kluyveromyces marxianus’, *Journal of Biotechnology*, 131(1), pp. 20–26. doi: 10.1016/j.jbiotec.2007.05.027.
Saini, R. *et al.* (2020) ‘Advanced biofuel production and road to commercialization: An insight into bioconversion potential of Rhodosporidium sp.’, *Biomass and Bioenergy*, 132(November 2019), p. 105439. doi: 10.1016/j.biombioe.2019.105439.
Sambrook, J.; Fritsch, E. F.; Maniatis, T. (1989) *Molecular cloning: a laboratory manual*. 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
Sarria, S. *et al.* (2014) ‘Microbial synthesis of pinene’, *ACS Synthetic Biology*, 3(7), pp. 466–475. doi: 10.1021/sb4001382.
Schultz, J. C., 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), pp. 2103– 2109. doi: https://doi.org/10.1002/bit.27001%0A.

De Schutter, K. *et al.* (2009) 'Genome sequence of the recombinant protein production host Pichia pastoris', *Nature Biotechnology*, 27(6), pp. 561–566. doi: 10.1038/nbt.1544.

Schwartz, C. *et al.* (2018) 'Validating genome-wide CRISPR-Cas9 function in the non-conventional yeast Yarrowia lipolytica', *bioRxiv preprint*. doi: http://dx.doi.org/10.1101/358630.

Shi, T. Q. *et al.* (2018) 'Advancing metabolic engineering of Yarrowia lipolytica using the CRISPR/Cas system', *Applied Microbiology and Biotechnology*, 102(22), pp. 9541–9548. doi: 10.1007/s00253-018-9366-x.

Shui, W. *et al.* (2015) 'Understanding the mechanism of thermotolerance distinct from heat shock response through proteomic analysis of industrial strains of saccharomyces cerevisiae', *Molecular and Cellular Proteomics*, 14(7), pp. 1885–1897. doi: 10.1074/mcp.M114.045781.

Soccol, C. R. *et al.* (2017) 'Pilot scale biodiesel production from microbial oil of Rhodosporidium toruloides DEBB 5533 using sugarcane juice: Performance in diesel engine and preliminary economic study', *Bioresource Technology*, 223, pp. 259–268. doi: 10.1016/j.biortech.2016.10.055.

Song, W. *et al.* (2016) 'Can terminators be used as insulators into yeast synthetic gene circuits', *Journal of Biological Engineering*, 10(1), pp. 1–13. doi: 10.1186/s13036-016-0040-5.

Sorek, N. *et al.* (2014) 'The implications of lignocellulosic biomass chemical composition for the production of advanced biofuels', *BioScience*, 64(3), pp. 192–201. doi: 10.1093/biosci/bit037.

Spohner, S. C. *et al.* (2016) 'Kluyveromyces lactis: An emerging tool in biotechnology', *Journal of Biotechnology*, 222, pp. 104–116. doi: 10.1016/j.jbiotec.2016.02.023.

Stanley, D. *et al.* (2010) 'The ethanol stress response and ethanol tolerance of Saccharomyces cerevisiae', *Journal of Applied Microbiology*, 109(1), pp. 13–24. doi: 10.1111/j.1365-2672.2009.04657.x.

Stephanopoulos, G. (2012) 'Synthetic biology and metabolic engineering', *ACS Synthetic Biology*, 1(11), pp. 514–525. doi: 10.1021/sb300094q.

Tan, E. C. D. and Lamers, P. (2021) 'Circular Bioeconomy Concepts—A Perspective', *Frontiers in Sustainability*, 2(July), pp. 1–8. doi: 10.3389/frsus.2021.701509.

Tanasova, M. and Sturla, S. J. (2012) 'Chemistry and Biology of Acylfulvenes: Sesquiterpene-Derived Antitumor Agents', *Chemical Reviews*, 112(6), pp. 3578–3610. doi: 10.1021/cr2001367.

The Ellen Macarthur Foundation (2022) *Let's build a circular economy*. Available at: https://ellenmacarthurfoundation.org/.

Tiukova, I. A. *et al.* (2019) 'Genome-scale model of Rhodotorula toruloides metabolism', *Biotechnology and Bioengineering*, 116(12), pp. 3396–3408. doi: 10.1002/bit.27162.

Tominaga, M., Kondo, A. and Ishii, J. (2022) 'Engineering of Synthetic Transcriptional Switches in Yeast', *Life*, 12(4). doi: 10.3390/life12040557.

Tremblay, B. (2020) 'universalmotif: Import, Modify, and Export Motifs with R.', *Bioconductor*, R package. Available at: https://bioconductor.org/packages/universalmotif/.

Tsai, Y. Y. *et al.* (2017) 'Development of a sufficient and effective procedure for transformation of an oleaginous yeast, Rhodosporidium toruloides DMKU3-TK16', *Current Genetics*, 63(2), p. 373. doi: 10.1007/s00294-017-0678-7.

Verghese, J. *et al.* (2012) 'Biology of the Heat Shock Response and Protein Chaperones: Budding Yeast (Saccharomyces cerevisiae) as a Model System', *Microbiology and Molecular Biology Reviews*, 76(2), pp. 115–158. doi: 10.1128/mmbr.05018-11.

Vogl, T., Gebbie, L., Palfreyman, Robin W., *et al.* (2018) 'Effect of plasmid design and type of integration event on recombinant protein expression in Pichia pastoris', *Applied and Environmental Microbiology*, 84(6). doi: 10.1128/AEM.02712-17.

Vogl, T., Gebbie, L., Palfreyman, Robin W, *et al.* (2018) 'Effect of plasmid design and type of integration event on recombinant protein expression in Pichia pastoris', *Applied and Environmental Microbiology*, 84(6). doi: 10.1128/AEM.02712-17.

Vogl, T., Kickenweiz, T., *et al.* (2018) 'Engineered bidirectional promoters enable rapid multi-gene co-expression optimization', *Nature Communications*, 9(1), p. 3589. doi: 10.1038/s41467-018-05915-w.

Vogl, T. and Glieder, A. (2013) 'Regulation of Pichia pastoris promoters and its

consequences for protein production', *New Biotechnology*, 30(4), pp. 385–404. doi: 10.1016/j.nbt.2012.11.010.
Vogl, T., Hartner, F. S. and Glieder, A. (2013) 'New opportunities by synthetic biology for biopharmaceutical production in Pichia pastoris', *Current Opinion in Biotechnology*, 24(6), pp. 1094–1101. doi: 10.1016/j.copbio.2013.02.024.
Wagle, Aditi *et al.* (2022) 'Multi-stage pre-treatment of lignocellulosic biomass for multi-product biorefinery: A review', *Sustainable Energy Technologies and Assessments*, 49(September 2021), p. 101702. doi: 10.1016/j.seta.2021.101702.
Wagner, James M and Alper, H. S. (2016) 'Synthetic biology and molecular genetics in non-conventional yeasts: Current tools and future advances', *Fungal Genetics and Biology*, 89, pp. 126–136. doi: 10.1016/j.fgb.2015.12.001.
Wagner, James M. and Alper, H. S. (2016) 'Synthetic biology and molecular genetics in non-conventional yeasts: Current tools and future advances', *Fungal Genetics and Biology*, 89, pp. 126–136. doi: 10.1016/j.fgb.2015.12.001.
Wang, Q., Quan, S. and Xiao, H. (2019) 'Towards efficient terpenoid biosynthesis: manipulating IPP and DMAPP supply', *Bioresources and Bioprocessing*, 6(1). doi: 10.1186/s40643-019-0242-z.
Wang, Y. *et al.* (2016) 'Cloning and evaluation of different constitutive promoters in the oleaginous yeast Rhodosporidium toruloides', *Yeast*, 33(3), pp. 99–106.
Wang, Y. *et al.* (2018) 'Systems analysis of phosphate-limitation -induced lipid accumulation by the oleaginous yeast Rhodosporidium toruloides', *Biotechnology for Biofuels*, pp. 1–15. doi: 10.1186/s13068-018-1134-8.
Weber, E. *et al.* (2011) 'A modular cloning system for standardized assembly of multigene constructs.', *PloS one*, 6(2), p. e16765. doi: 10.1371/journal.pone.0016765.
Wehrs, M. *et al.* (2019) 'Engineering Robust Production Microbes for Large-Scale Cultivation', *Trends in Microbiology*, xx, pp. 1–14. doi: 10.1016/J.TIM.2019.01.006.
Wei, L. J. *et al.* (2021) 'Biosynthesis of α-Pinene by Genetically Engineered Yarrowia lipolytica from Low-Cost Renewable Feedstocks', *Journal of Agricultural and Food Chemistry*, 69(1), pp. 275–285. doi: 10.1021/acs.jafc.0c06504.
Wen, Z. *et al.* (2020) 'Rhodosporidium toruloides - A potential red yeast chassis for lipids and beyond', *FEMS Yeast Research*, 20(5), pp. 1–12. doi:

10.1093/femsyr/foaa038.
Wickham, H. (2016) ‘ggplot2: Elegant Graphics for Data Analysis.’, *Springer-Verlag New York*.
Woodroffe, J. D. *et al.* (2021) ‘Synthesis and fuel properties of high-energy density cyclopropanated monoterpenes’, *Fuel Processing Technology*, 222(April), p. 106952. doi: 10.1016/j.fuproc.2021.106952.
World Economic Forum (2020) *Why the world needs a ‘circular bioeconomy’ - for jobs, biodiversity and prosperity*. Available at: https://www.weforum.org/agenda/2020/10/circular-bioeconomy-nature-reset/.
Wu, C.-C. *et al.* (2020) ‘Ethanol and H2 O2 stresses enhance lipid production in an oleaginous Rhodotorula toruloides thermotolerant mutant L1-1’, *FEMS Yeast Research*, 20(4).
Yaashikaa, P. R., Senthil Kumar, P. and Varjani, S. (2022) ‘Valorization of agro-industrial wastes for biorefinery process and circular bioeconomy: A critical review’, *Bioresource Technology*, 343(August 2021), p. 126126. doi: 10.1016/j.biortech.2021.126126.
Yaegashi, J., Kirby, J., Ito, M., Sun, J., Dutta, T., Mirsiaghi, M., Sundstrom, Eric R, *et al.* (2017) ‘Rhodosporidium toruloides: A new platform organism for conversion of lignocellulose into terpene biofuels and bioproducts’, *Biotechnology for Biofuels*, 10(1), pp. 1–13. doi: 10.1186/s13068-017-0927-5.
Yaegashi, J., Kirby, J., Ito, M., Sun, J., Dutta, T., Mirsiaghi, M., Sundstrom, Eric R., *et al.* (2017) ‘Rhodosporidium toruloides: A new platform organism for conversion of lignocellulose into terpene biofuels and bioproducts’, *Biotechnology for Biofuels*, 10(1), pp. 1–13. doi: 10.1186/s13068-017-0927-5.
Yang, C. *et al.* (2015) ‘Characterizing yeast promoters used in Kluyveromyces marxianus’, *World Journal of Microbiology and Biotechnology*, 31(10), pp. 1641–1646. doi: 10.1007/s11274-015-1899-x.
Yang, Z., Edwards, H. and Xu, P. (2020) ‘CRISPR-Cas12a/Cpf1-assisted precise, efficient and multiplexed genome-editing in Yarrowia lipolytica’, *Metabolic Engineering Communications*, 10(November 2019), p. e00112. doi: 10.1016/j.mec.2019.e00112.
Yuan, J. and Ching, C. B. (2016) ‘Mitochondrial acetyl-CoA utilization pathway for

terpenoid productions', *Metabolic Engineering*, 38, pp. 303–309. doi: 10.1016/j.ymben.2016.07.008.

Zhang, G. *et al.* (2022) 'Metabolic engineering of Yarrowia lipolytica for terpenoids production: advances and perspectives', *Critical Reviews in Biotechnology*, 42(4), pp. 618–633. doi: 10.1080/07388551.2021.1947183.

Zhang, S. *et al.* (2016) 'Metabolic engineering of the oleaginous yeast Rhodosporidium toruloides IFO0880 for lipid overproduction during high-density fermentation', *Applied Microbiology and Biotechnology*, 100(21), pp. 9393–9405. doi: 10.1007/s00253-016-7815-y.

Zhou, P. *et al.* (2020) 'Improved linalool production in Saccharomyces cerevisiae by combining directed evolution of linalool synthase and overexpression of the complete mevalonate pathway', *Biochemical Engineering Journal*, 161(May). doi: 10.1016/j.bej.2020.107655.

Zhuang, X. *et al.* (2019) 'Monoterpene production by the carotenogenic yeast Rhodosporidium toruloides', *Microbial Cell Factories*, 18(1), pp. 1–15. doi: 10.1186/s12934-019-1099-8.