



UNIVERSIDADE ESTADUAL DE CAMPINAS
Instituto de Biologia

FELIPE EDUARDO CIAMPONI

CHARACTERIZATION OF THE GENO-TRANSCRIPTOMIC PROFILE OF AN INDUSTRIAL
SACCHAROMYCES CEREVISIAE STRAIN IN RESPONSE TO STRESS INDUCED BY
PARA-COUMARIC ACID

CARACTERIZAÇÃO DO PERFIL GENO-TRANSCRIPTÔMICO DE UMA LINHAGEM
INDUSTRIAL DE *SACCHAROMYCES CEREVISIAE* EM RESPOSTA A ESTRESSE
INDUZIDO POR ÁCIDO PARA-CUMÁRICO

Campinas
2022

FELIPE EDUARDO CIAMPONI

CHARACTERIZATION OF THE GENO-TRANSCRIPTOMIC PROFILE OF AN INDUSTRIAL
SACCHAROMYCES CEREVISIAE STRAIN IN RESPONSE TO STRESS INDUCED BY
PARA-COUMARIC ACID

CARACTERIZAÇÃO DO PERFIL GENO-TRANSCRIPTÔMICO DE UMA LINHAGEM
INDUSTRIAL DE SACCHAROMYCES CEREVISIAE EM RESPOSTA A STRESS INDUZIDO
POR ÁCIDO PARA-CUMÁRICO

*Thesis presented to the Biology Institute of the
University of Campinas in partial fulfillment of
the requirements for the degree of Doctor in
Genetics and Molecular Biology in the area of
Bioinformatics*

*Tese apresentada ao Instituto de Biologia da
Universidade Estadual de Campinas como
parte dos requisitos exigidos para a obtenção
do Título de Doutor em Genética e Biologia
Molecular na área de Bioinformática*

Advisor: Dr. Marcelo Mendes Brandão

ESTE ARQUIVO DIGITAL CORRESPONDE
À VERSÃO FINAL DA TESE DEFENDIDA
PELO ALUNO FELIPE EDUARDO
CIAMPONI E ORIENTADA PELO PROF. DR.
MARCELO MENDES BRANDÃO.

Campinas
2022

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Biologia
Mara Janaina de Oliveira - CRB 8/6972

C481c Ciamponi, Felipe Eduardo, 1991-
Characterization of the geno-transcriptomic profile of an industrial
Saccharomyces cerevisiae strain in response to stress induced by para-
coumaric acid / Felipe Eduardo Ciamponi. – Campinas, SP : [s.n.], 2022.

Orientador: Marcelo Mendes Brandão.

Tese (doutorado) – Universidade Estadual de Campinas, Instituto de
Biologia.

1. *Saccharomyces cerevisiae*. 2. Transcriptoma. 3. Genômica comparativa.
4. Banco de dados. 5. Redes complexas. I. Brandão, Marcelo Mendes, 1974-
II. Basso, Thiago Olitta. III. Universidade Estadual de Campinas. Instituto de
Biologia. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Caracterização do perfil geno-transcriptômico de uma linhagem industrial de *Saccharomyces cerevisiae* em resposta a estresse induzido por ácido paracumárico

Palavras-chave em inglês:

Saccharomyces cerevisiae

Transcriptome

Comparative genomics

Databases

Complex networks

Área de concentração: Bioinformática

Titulação: Doutor em Genética e Biologia Molecular

Banca examinadora:

Marcelo Mendes Brandão [Orientador]

André Lima Damásio

Andreas Karoly Gombert

Wendel Batista da Silveira

Gabriela Felix Persinoti

Data de defesa: 19-05-2022

Programa de Pós-Graduação: Genética e Biologia Molecular

Identificação e informações acadêmicas do(a) aluno(a)

- ORCID do autor: <https://orcid.org/0000-0002-0076-882>

- Currículo Lattes do autor: <http://lattes.cnpq.br/2256041278569579>

Campinas, 19 de Maio de 2022.

COMISSÃO EXAMINADORA

Prof. Dr. Marcelo Mendes Brandão

Prof. Dr. André Ricardo de Lima Damásio

Profa. Dra. Gabriela Felix Persinoti

Prof. Dr.. Andreas Karoly Gombert

Prof. Dr. Wendel Batista da Silveira

Os membros da Comissão Examinadora acima assinaram a Ata de defesa, que se encontra no processo de vida acadêmica do aluno.

A Ata da defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa Genética e Biologia Molecular da Unidade Instituto de Biologia.

DEDICATORY

I dedicate this work to all my family, friends, colleagues, advisors and teachers that were part of my journey, both personal and academic, throughout the years. Your support and incentive were crucial to get me where I am today, this work would not exist without all of you.

“If I have seen further, it is by standing on the shoulders of giants.”

- Sir Isaac Newton

ACKNOWLEDGMENTS

I would like to thank my family and friends for all the support in this journey. Always by my side, crying and celebrating with me at every step. You helped me through a lot of bad times and also cheered with me at the good times.

To my supervisor, Dr. Marcelo Mendes Brandão, for the opportunities and patience during the 4 years I have spent in his lab. Your guidance and wisdom was invaluable, not only did you help me become a better scientist but also a better person. You are the role model of a scientist that I aspire to be.

To Prof. Dr. Thiago Olitta Basso, who was my co-advisor and provided valuable insights during several discussions. Your knowledge and contributions were crucial for making this project a reality.

To my lab colleagues, who became my second family during these years. With special thanks to Natália Faraj Murad, Murilo Meneghetti and Dielle Pierotti Procópio who worked directly in the projects presented here.

To Prof. Dr. Karina Lucas da Silva-Brandão, whose expertise on comparative genomics contributed immeasurably to this project.

To Prof. Dr. André Ricardo de Lima Damasio, Prof. Dr. Anderson Ferreira da Cunha and Prof. Dr. Marcus Bruno Soares Forte for the valuable feedback and ideas provided during my qualification exam.

To the State University of Campinas and the Center for Molecular Biology and Genetic Engineering for hosting this research project.

To the funding agencies CAPES, CNPq and FAPESP that supported this study via fellowships and research grants.

This work was carried out with the support of the Higher Education Personnel Improvement Coordination - Brazil (CAPES) - Financing Code 001

RESUMO

O etanol de segunda geração tem estado na vanguarda dos esforços de pesquisa em bioenergia na última década. Isso se deve aos potenciais benefícios do uso da biomassa lignocelulósica como fonte de energia, que superam os maiores custos associados à produção industrial desse tipo de combustível. No entanto, ainda existem alguns desafios que precisam ser superados para que essa tecnologia seja uma alternativa viável para a produção de biocombustíveis. Nesta tese, abordaremos duas dessas questões: a necessidade de cepas de *Saccharomyces cerevisiae* cada vez mais robustas e a falta de um banco de dados centralizado para genômica comparativa de cepas industriais.

No primeiro capítulo desta tese, apresentamos nosso modelo de rede multi-ômica que desenvolvemos para estudar os mecanismos de sobrevivência da SA-1, uma cepa de levedura industrial altamente resistente, quando exposta ao ácido p-cumárico (pCA). Este trabalho demonstra o uso de uma abordagem combinada de medição de metabólitos, perfil de transcriptoma e descoberta de variante genômica para construir um modelo integrado construído em um formato baseado em gráfico. Usando este modelo de rede, conseguimos determinar não apenas quais vias biológicas são alteradas durante a resposta ao pCA, mas também quais genes atuam como focos de interação, permitindo a descoberta de potenciais alvos de interesse para investigação biotecnológica.

No segundo bate-papo, apresentamos o INDYdb, um projeto de banco de dados desenvolvido por nosso grupo, cujo objetivo é criar uma plataforma especializada em linhagens de leveduras industriais para análise comparativa de genoma. Usando algoritmos avançados, clusters de computadores poderosos e uma interface amigável, conseguimos desenvolver um pipeline capaz de gerar anotações confiáveis de maneira consistente e escalável. Ao aplicar esta abordagem a 26 cepas de leveduras diferentes, fomos capazes de demonstrar as aplicações do INDYdb em três cenários diferentes: Reconstruindo a história evolutiva de *S. cerevisiae* por meio de genômica comparativa, identificando as principais características associadas a um gene alvo de interesse e gerando um hipótese específica baseada em uma busca exploratória do banco de dados a partir de uma questão biológica.

Esses dois trabalhos estão no contexto do projeto colaborativo FAPESP-BBRSC (2015/50612-8) intitulado "Uma abordagem integrada para explorar um novo paradigma para a produção de biocombustíveis a partir de matérias-primas lignocelulósicas". Trata-se de um esforço conjunto entre 10 grupos de trabalho diferentes espalhados por 3 universidades diferentes (Unicamp e USP no Brasil e Bath na Inglaterra) e é um componente fundamental do Centro de Pesquisas da Biorrefinaria.

O principal objetivo do nosso projeto é obter novos insights sobre os aspectos biológicos do metabolismo de leveduras industriais, criar bases de conhecimento que possam ser compartilhadas com a comunidade de pesquisa em bioengenharia e lançar as bases para projetos de longo prazo realizados por outros grupos. Tomados em conjunto, os projetos discutidos neste artigo atingem essa tríade de objetivos. Prevemos que, combinando abordagens de biologia computacional de última geração com uma extensa rede colaborativa, os dados apresentados aqui serão uma importante fonte de informações para pesquisadores que trabalham em bioengenharia de *Saccharomyces cerevisiae* em todo o mundo.

ABSTRACT

Second-generation ethanol has been at the forefront of bioenergy research efforts for the past decade. This is due to the potential benefits of using lignocellulosic biomass as an energy source, which outweigh the higher costs associated with industrial production of this type of fuel. However, there are still some challenges that need to be overcome in order to make this technology a viable alternative for biofuel production. In this thesis, we will address two of these issues: the need for increasingly robust *Saccharomyces cerevisiae* strains and the lack of a centralized database for comparative genomics of industrial strains.

In the first chapter of this thesis, we present our multi-omics network model that we developed to study the survival mechanisms of SA-1, a highly resistant industrial yeast strain, when exposed to p-coumaric acid (pCA). This work demonstrates the use of a combined approach of metabolite measurement, transcriptome profiling, and genomic variant discovery to build an integrated model constructed in a graph-based format. Using this network model, we were able to determine not only which biological pathways are altered during the response to pCA, but also which genes act as foci of interaction, allowing the discovery of potential targets of interest for biotechnological investigation.

In the second chapter, we present INDYdb, a database project developed by our group, whose goal is to create a platform specialized in industrial yeast strains for comparative genome analysis. By using advanced algorithms, powerful computer clusters, and a user-friendly interface, we were able to develop a pipeline capable of generating trustworthy annotations in a consistent and scalable manner. By applying this approach to 26 different yeast strains, we were able to demonstrate the applications of INDYdb in three different scenarios: Reconstructing the evolutionary history of *S. cerevisiae* through comparative genomics, identifying key features associated with a target gene of interest, and generating a specific hypothesis based on an exploratory search of the database starting from a biological question.

These two works are in the context of the FAPESP-BBRSC collaborative project (2015/50612-8) entitled "An integrated approach to explore a novel paradigm

for biofuel production from lignocellulosic feedstocks". This is a joint effort between 10 different working groups spread across 3 different universities (Unicamp and USP in Brazil and Bath in England) and is a key component of the Biorefinery Research Center. The main goal of our project is to gain new insights into the biological aspects of industrial yeast metabolism, create knowledge bases that can be shared with the bioengineering research community, and lay the groundwork for long-term projects undertaken by other groups. Taken together, the projects discussed in this paper achieve this trifecta of goals. We anticipate that by combining state-of-the-art computational biology approaches with an extensive collaborative network, the data presented here will be an important source of information for researchers working on *Saccharomyces cerevisiae* bioengineering around the world.

ABBREVIATIONS AND ACRONYMS

a.k.a. – “also known as”

API - Application Programming Interface

BH - Benjamini-Hochberg

BUSCO – Benchmark universal single copy orthologs

CDS – Coding DNA sequence

CSV – Comma-separated values

DeNSAS – *De Novo* Sequence Annotation System

DOI – Digital Object identifier

e.g. - “*exempli gratia*” or “for example”

FC – Fold change

FDR – False discovery rate

GFF – General feature format

GTF – Gene transfer format

HPC – High performance computing

HPLC – High performance liquid chromatography

i.e. – “*id est*” or “that is”

INDYdb – Industrial yeasts database

JSON – JavaScript Object Notation

LO – LiftOff

ORF – Open reading Frame

pCA – para-Coumaric acid

pVal – p-Value

REST - Representational state transfer

RNA-seq – High-throughput RNA sequencing

SGD – *Saccharomyces* genome database

SNP – Single nucleotide polymorphism

SNV – Single nucleotide variant (site)

SV – Short variant

TABLE OF CONTENTS

INTRODUCTION	14
CHAPTER 1 - Multi-omics network model reveals key genes associated to p-coumaric acid stress response in an industrial yeast strain	19
ABSTRACT	19
MATERIAL AND METHODS	19
Yeast strain and media	19
Analysis of extracellular metabolites	20
RNA extraction and sequencing	21
Gene expression analysis and functional characterization	22
Co-expressed gene cluster detection and hub gene identification	22
Short variant discovery	23
Multi-omics network assembly	23
Comparative genomics of SA-1 genes	24
RESULTS	26
Metabolic changes in SA1 yeasts upon pCA exposure	26
Differential gene expression between treated and control samples	27
Functional characterization of co-expressed gene clusters	31
Prediction of genomic short-variants based on RNA-seq	34
Assembly of a multi-omics network model for pCA response	37
Identification of distinguishable genomic features for SA-1 yeast strain	40
DISCUSSION	45
CONCLUSION	53
CHAPTER 2 - INDYdb: An integrative comparative genomics database for industrial yeast strains	55
ABSTRACT	55
MATERIAL AND METHODS	55
Yeast strain and genome assembly selection	56
Structural gene annotation	56
Functional gene annotation	57

Annotation benchmark	57
Data integration and Database assembly	58
Web access and API Interface	58
Comparative genomics and character tracing	58
RESULTS AND DISCUSSION	59
Annotation pipeline	59
Benchmarking annotation pipeline	62
Current strains of INDYdb	65
Comparative genomics of INDYdb's strains	69
Investigating specific gene of interest	71
Prospection of potential targets based on phenotype	74
CONCLUSION	76
FINAL REMARKS	77
REFERENCES	79
ANNEX A - Bioethics and Biosafety Statement	96
ANNEX B - Copyright Statement	97

INTRODUCTION

CHAPTER 1 - Multi-omics network model reveals key genes associated to p-coumaric acid stress response in an industrial yeast strain

Expansion of the global lignocellulosic ethanol production is heading towards second-generation ethanol (2G) biorefineries. Although 2G ethanol is still more expensive than first-generation ethanol (1G), current production costs for 2G being up to 50% higher than 1G (JUNQUEIRA et al., 2017; PETERSEN et al., 2021), recent advances in biofuel technology suggest that 2G ethanol will be more cost-efficient in the long run, with some of the more optimistic scenario placing the turning point for this technology in the year 2025 (JUNQUEIRA et al., 2017; RAJ et al., 2022; TAPIA CARPIO; SIMONE DE SOUZA, 2019). For commodities like ethanol, even small changes in production costs can have a big impact on the supply chain. Reducing operating costs by a few cents can result in savings of millions of dollars per year. (KOHLER, 2019; MCALOON; TAYLOR; YEE, 2000; MIZIK, 2020).

During 2G ethanol production, before the biomass being placed in fermentation vats, the lignocellulosic feedstock undergoes pre-processing to unleash less complex sugars located in the cell wall to make these molecules available for the yeasts (TUMULURU, 2018). However this process also releases several toxic compounds in the medium, and, as a result, it requires microorganisms with increasing resistance to inhibitors generated during pretreatment processes (MAURYA; SINGLA; NEGI, 2015; SINDHU; BINOD; PANDEY, 2016). Therefore, understanding how these inhibitory molecules affect the fermentative performance of *Saccharomyces cerevisiae* is essential to implement strategies to increase its robustness against adverse conditions in industrial fermentation (ALMEIDA et al., 2007; HAMELINCK; VAN HOOIJDONK; FAAIJ, 2005; SAMBUSITI et al., 2013) and contributing to its implementation as a stable platform for biofuel production.

Among these compounds, two classes of molecules – furans and organic acids – have active physiological impact on the growth rate and overall fermentation metabolism of *S. cerevisiae* (ALMEIDA et al., 2007; LARSSON et al., 1999, 2000;

NILSSON et al., 2005; RUSSELL, 1992; SAMBUSITI et al., 2013; TAHERZADEH et al., 2000). Additionally, phenolic compounds also inhibit the production of ethanol in anaerobic fermentation (SAMBUSITI et al., 2013). Although certain *S. cerevisiae* strains are resistant to these molecules, the molecular mechanism used by these yeasts to metabolize such inhibitors into less toxic compounds are complex, involving multiple regulatory processes and pathways (FAVARO; JANSEN; VAN ZYL, 2019). One of the major byproducts resulting from the sugarcane bagasse pretreatment in the production of second-generation bioethanol is pCA (pCA) (JÖNSSON; MARTÍN, 2016; REINOSO et al., 2018), which was found to be in concentrations of up to 2.0 g/kg (dry weight) of bagasse after pretreatment (BIAZI et al., 2022; VAN DER POL et al., 2015). This chemical usually inhibits the growth of *S. cerevisiae* and disrupts the production of ethanol (ADEBOYE; BETTIGA; OLSSON, 2017; BARANOWSKI et al., 1980; COLA et al., 2020; GU; ZHANG; BAO, 2014). Although some reports show that certain *S. cerevisiae* strains are capable of surviving high concentrations of pCA, or even being engineered to serve as templates for production of this compound (BORJA et al., 2019; LIU et al., 2019b), the same cannot be said for strains currently in use in the 2G bioethanol industry. These strains have a significantly different genomic makeup, in the form of nucleotide variation and structural rearrangements, caused by the intense selection process that this particular group of *S. cerevisiae* was subjected to (JACOBUS et al., 2021; WOHLBACH et al., 2014; ZHANG et al., 2015, 2016), and appear to be more susceptible to the inhibitory effects of these compounds (COLA et al., 2020; MORENO et al., 2019; VAN DER POL et al., 2014); Moreover, pCA is insoluble in water, but easily reacts and solubilizes in ethanol, posing a significant challenge for any type of industrial-scale fermentation process (SALAMEH et al., 2008).

Recently, a Brazilian industrial strain used in the bioethanol industry called SA-1 was shown to be highly resistant to several lignocellulosic inhibitors, being capable of maintaining 70% of its normal growth rate even when exposed to 7 mM of pCA, a feat that was not observed even in other industrial strains, such as the case of JAY270, a haploid derivative of PE-2 which is one of the most widespread strains currently in use in the Brazilian bioethanol industry (COLA et al., 2020; DE MELLO et al., 2019; NAGAMATSU et al., 2019). However, the molecular characterization of

specific survival mechanisms used by SA-1 to survive in such conditions has not yet been described. Considering that the response of *S. cerevisiae* gene expression to environmental conditions is a powerful tool for identifying targets associated with increased ethanol production and survivability and has been used to direct bioengineering efforts toward a desired phenotype by altering the transcriptional machinery of these organisms (ALPER et al., 2006; LIN; ZHANG; WANG, 2013; TECHAPARIN; THANONKEO; KLANRIT, 2017). The use of data obtained for differentially expressed genes in combination with systems biology approaches also allowed researchers to identify the metabolic pathways that are affected (activated or repressed) under specific conditions (FENG; ZHAO, 2013; HANCOCK; TAKIGAWA; MAMITSUKA, 2010; HERNÁNDEZ-ELVIRA; SUNNERHAGEN, 2022; TARCA et al., 2009).

Our study focuses on characterizing the SA-1 strain profile when exposed to high concentrations of pCA under continuous fermentation conditions in a controlled bioreactor environment, using a combination of metabolite analysis, transcriptomics and genomics in an integrative and systemic multi-omics analysis to elucidate the underlying mechanisms by which this particular strain is capable of thriving even when exposed to such inhibitors. Our main objectives are to not only characterize the molecular aspects of the response, but also identify key genes associated with the response to such inhibitors, allowing a deeper understanding of these functions. By using these findings as a framework for future bioengineering efforts, be it in the form of gene models or pathways of interest, we expect our work to provide valuable insights towards the development of more robust industrial strains that are capable of increased survival rates when exposed to the adversary conditions present in industrial fermentation vats, assisting in the reduction overall production costs of 2G ethanol production and establishing this platform as a stable source for biofuel production.

CHAPTER 2 - INDYdb: An integrative comparative genomics database for industrial yeast strains

Commonly called baker's yeast, the *Saccharomyces cerevisiae* is one of the most important organisms in use for industries, applied sciences and basic

research (HANSON, 2018; PARAPOULI et al., 2020). This unicellular organism has been used by humanity for millennia, from the first time a fermented beverage was brewed to the latest development on second generation ethanol, *S. cerevisiae* has been in the center of multiple fields that range from food and beverage industries to biofuel production, pharmaceutical development, cell factories and an assortment of other areas (LIU et al., 2017; POMPON, 1999; WALKER; PRETORIUS, 2018).

All of these fields share the peculiarity of being highly competitive, either via publications or patents, with research groups all around the world racing to achieve the next breakthrough. Recent studies show that the swiftness by which relevant data can be gathered from datasets plays a much more prominent role than the amount of data available in pushing technological development (AL NAQBIA et al., 2020), with innovation being one of the most important factors for determining economic growth for public, private or public-private partnership enterprises (CARBONARA; PELLEGRINO, 2020; KHAN et al., 2020; PRADHAN et al., 2020; SURYA et al., 2021).

Databases are one of the key drivers of innovation for scientists working with bioinformatics, offering researchers ways to access data relevant to their field of study by consulting publicly available datasets or information gathered from the literature (BAXEVANIS; BATEMAN, 2015). However, these databases can present their data in vastly different formats that vary according to the purpose of the repository as well as the method by which they share the information (CHEN; HUANG; WU, 2017; RIGDEN; FERNÁNDEZ, 2019). The increasing number of publications and the amount of data generated and deposited in such databases has made the task of extracting meaningful biological information from these sources an ever-increasing challenge that researchers must overcome on a daily basis. (BORNMANN; HAUNSCHILD; MUTZ, 2021; IMKER, 2018). This issue was further pushed to the limit with the increased focus on systems biology and multi-omics studies (JAISWAL et al., 2020; KIM et al., 2017; SUBRAMANIAN et al., 2020; YAHYA et al., 2021).

In order to mitigate this, the bioinformatics community has been using, for the past decade, several Big Data analysis and Data Mining strategies to improve not only the quality of data extracted from databases, but also its biological significance

(GUPTA; CHANDRA, 2020; LIN, 2017; PAL et al., 2020; SINGH; SINGH, 2020). This led to the growth of integrative databases, which compile data from public datasets and provide researchers with an all-in-one package that is both user-friendly and contains information that would otherwise be splintered across multiple sources (VILLALBA; MATTE, 2021). These types of databases can be found in use for multiple fields: Gene expression (BONO, 2020), phylogenetics (CHOROSTECKI et al., 2021), microbiome (CHEN et al., 2020), metagenomics (TU et al., 2019) and even microRNA target predictions (TOKAR et al., 2018).

Unfortunately, when we look at the current state of integrative databases focused on industrial yeast strains the overall panorama is bleak. Saccharomyces Genome Database (CHERRY et al., 2012), one of the most important resources for functional annotation on *S. cerevisiae*, only has information on a little more than 50 strains. Other major databases of yeast biological functions, such as KEGG (KANEHISA et al., 2017), Yeast BioCyc (KARP et al., 2018) and Yeast Metabolome Database (RAMIREZ-GAONA et al., 2017) focus on information derived only from the S288C reference strain. With over 1000 *S. cerevisiae* genome assemblies currently available on GenBank (SAYERS et al., 2019) and the increase in relevance of comparative genomics studies for both industrial and laboratory applications (BORELLI et al., 2019; GALLONE et al., 2016, 2018; JACOBUS et al., 2021; MARSIT et al., 2017), the need for an integrative database that focuses on non-reference strains is clear.

In order to address this issue, we present INDYdb, an integrative database for comparative genomics of yeast strains commonly used in industrial and laboratory applications. Our database provides structural and functional gene annotation for each strain, we also provide comprehensive comparisons across strains such as: Multiple sequence alignment, variation calls and phylogenetic analysis. By creating a centralized source of information that allows comparison between strains currently in use in multiple fields, both in industry and in basic research, we intended to provide researchers around the globe with the means to quickly and easily analyze a multitude of characteristics derived from yeast comparative genomics, bolstering their efforts and expedite biotechnological innovation.

CHAPTER 1 - Multi-omics network model reveals key genes associated to p-coumaric acid stress response in an industrial yeast strain

ABSTRACT

The production of ethanol from lignocellulosic sources presents increasingly difficult issues for the global biofuel scenario, leading to increased production cost of current second-generation (2G) ethanol when compared to first-generation (1G) plants. Among the setbacks encountered in industrial processes, the presence of chemical inhibitors from pre-treatment processes severely hinder the potential of yeasts in producing ethanol at peak efficiency. However, some industrial yeast strains have, either naturally or artificially, higher tolerance levels to these compounds. Such is the case of SA-1, a Brazilian industrial strain that has shown high resistance to inhibitors produced by the pre-treatment of cellulosic complexes. Our study focuses on the characterization of the transcriptomic and physiological impact of an inhibitor of this type, p-Coumaric acid (pCA), on this strain under chemostat cultivation via RNAseq and HPLC data. We show that, when exposed to pCA, SA-1 yeasts tend to increase ethanol production while reducing overall biomass yield, as opposed to pCA-susceptible strains that tend to reduce their fermentation efficiency when exposed to this compound, suggesting increased metabolic activity associated to mitochondrial and peroxisomal processes. Transcriptomic analysis also revealed a plethora of differentially expressed genes located in co-expressed clusters that are associated with changes in biological pathways linked to biosynthetic and energetical processes. Furthermore, we also identified 20 genes that act as interaction hubs for these clusters, while also having association with altered pathways and changes in metabolic outputs, potentially leading to the discovery of novel targets for genetic engineering towards a more robust industrial yeast strain.

MATERIAL AND METHODS

Yeast strain and media

The strain investigated in this study, SA-1, is a derived industrial strain (MATa/MAT α) isolated and distributed by Fermentec (Piracicaba, Brazil). Inoculum cultures were prepared from glycerol stocks stored at -80°C on a defined medium

(LUTTIK et al., 2000; VERDUYN et al., 1992), whose composition (in g.L⁻¹) is described as follows: NH₂CONH₂ (urea), 2.3; KH₂PO₄, 3.0; K₂SO₄, 6.6; MgSO₄·7H₂O, 0.5; 1 mL L⁻¹ trace element solution, 1 mL L⁻¹ vitamin solution, and 20 g L⁻¹ glucose. Cultures were grown overnight at 30°C in a rotary shaker at 200 rpm. Chemostat cultivation with pCA with SA-1 *S. cerevisiae* strains was performed in a 2.0-liter water jacket model Labfors 5 (Infors AG, Switzerland) with 1.0 liter working volume, which was kept constant by a mechanical drain and a peristaltic pump. Throughout cultivation, both the culture vessel (0.5 l min⁻¹) and the medium vessel (flow rate not measured) were purged with nitrogen gas to maintain anaerobic conditions. The circulation frequency was set at 800 rpm, the temperature was controlled at 30 °C, and the pH was adjusted to 5.0 using a controlled 2 M KOH solution. Pre-cultures for batch bioreactor cultivations were grown overnight in an orbital shaker at 30 °C and 200 rpm in 500-mL shake flasks containing 100 ml of the defined medium with 20 g L⁻¹ starting glucose. The medium had the same composition as the preculture, except that Tween 80 and ergosterol were added at a final concentration of 0.01 g L⁻¹ and 0.42 g L⁻¹, respectively, to allow anaerobic growth. The batch phase was terminated after glucose depletion (monitored by a sharp drop in CO₂ concentration in the exhaust gas), whereupon cultivation switched to continuous mode with addition of fresh medium supplemented or not supplemented with 7 mM pCA. The dilution rate was set at 0.1 h⁻¹ and the cultivation was assumed to be in steady state when the dry weight of the culture and the specific carbon dioxide production rate varied by less than 2% for two volume changes during at least five residence times.

The chemostat system was chosen due to its characteristics of maintaining physiological conditions in constant values among experiments, which is important when trying to isolate transcriptomic alterations that arise in response to a singular input (in our case, the presence of pCA), eliminating the effects of growth rates and other stochastic perturbations which may arise due to environmental conditions (REGENBERG et al., 2006).

Analysis of extracellular metabolites

Cell dry mass concentration was determined by gravimetric method (OLSSON; NIELSEN, 1997). Extracellular metabolite samples from the chemostat cultures were filtered through 0.2 µm syringe filters. Concentrations of residual carbon,

ethanol, glycerol, and organic acids were quantified by high-performance liquid chromatography (HPLC) (DELLA-BIANCA et al., 2014), using a Prominence HPLC model (Shimadzu Corporation, Japan) and an HPX-87H analytical column (Bio-Rad Laboratories, USA) at 60 °C with 5 mM H₂SO₄ as mobile phase at 0.6 mL min⁻¹. Ethanol concentrations were corrected for evaporation (MEDINA et al., 2010) and pCA was quantified (KAMMERER et al., 2004) by using an HPLC and a C18 analytical column (Supelco Inc. model Waters Spherisorb ODS - 25 µm, 250 mm x 4.6 mm) at 30 °C with 2% (v/v) acetic acid in ionized water (eluent A) and acetic acid 0.5 % in ionized water and acetonitrile (50:50, v/v; eluent B) as mobile phase at 1.0 mL min⁻¹ using a gradient program: from 10 to 15 % B (10 min), 15 % B isocratic (3 min), 15 to 25 % B (7 min), 25 to 55 % B (30 min), 55 to 100 % B (1 min), 100 % B isocratic (5 min), from 100 to 10 % B (0.1 min). The total run time was 60 min, with a flow rate of 1.0 mL min⁻¹ and an oven temperature of 30 °C. The injection volume for all samples was 10 µL. Monitoring was performed using a Shimadzu UV detector at wavelengths of 280 nm and 320 nm. Concentrations of compounds were calculated from calibration curves obtained from standard solutions.

RNA extraction and sequencing

RNA extraction was performed using Direct-zol™ RNA MiniPrep kit (Zymo Research catalog no. R2051) following manufacturer's instructions. RNA samples were sequenced using BGISEQ-500, with each library generating approximately 24M paired-end reads of 100bp. Raw RNA-seq reads were filtered to remove adapter contamination and low-quality reads, with Trimmomatic (BOLGER; LOHSE; USADEL, 2014). Each sample was aligned against the R64-1-1 version of the *S. cerevisiae* reference genome, which is based on the S288C strain, with STAR v2.7.0 (DOBIN et al., 2013), using “—sjdbGTFfile,” “--quantMode GeneCounts”, “--twopassMode Basic” and the ENCODE guidelines for best practices of eukaryotic RNASeq (ENCODE, 2016) as additional parameters. The corresponding gene annotation files and variant call files were also obtained for the same assembly. All genome data was obtained from Ensembl Fungi release 48 (HOWE et al., 2020).

Gene expression analysis and functional characterization

Differential gene expression was assessed by edgeR v.3.3 (ROBINSON; MCCARTHY; SMYTH, 2010), using $FDR \leq 0.01$ and $|\log_2(\text{FoldChange})| \geq 0.5$ as cutoffs for statistical significance. Gene expression in $\log_2(\text{CPM})$ scale was used to perform principal component analysis (PCA) and replicate similarity assessment to check the significance of biological duplicates. All downstream functional enrichment analyses were done using STRINGdb v.11 (SZKLARCZYK et al., 2019), and pathway perturbation analysis with Pathview API and GAGE v.2.38 (LUO et al., 2009, 2017). In both cases an FDR cutoff of 0.01 was applied using the coding genome as background, and the Fold enrichment value was calculated based on the number of observed genes in comparison to the number of expected hits.

Co-expressed gene cluster detection and hub gene identification

Co-expressed gene clusters were identified using an adaptation of the kNN-enhance method, which intensifies an existing network with node attributes (JIA et al., 2017). The total protein-protein interaction (ppi) network from STRINGdb (v11) was converted into an undirected graph, where each node is a protein and the edges represent known interactions between them – only interactions with a total ppi_score ≥ 0.7 (high confidence) were considered for downstream analysis. Each node metadata information was enhanced with an extra attribute corresponding to the $\log_2(\text{foldChange})$ value of that protein, and for each pair of vertices connected by an edge a second score was used (called foldChange_score). This metric was calculated by $1 - \text{norm}\left(\left(X_i - X_j\right)_2\right)$ where X_i is the foldChange for vertex 1, and X_j is the foldChange for vertex 2, and $\text{norm}\left(\left(X_i - X_j\right)_2\right)$ is the normalized Euclidean distance between X_i and X_j . Thus, values for foldChange_score varied from 1 (identical foldChange scores) to 0 (the largest foldChange difference between two nodes in the network).

Final edge weight scores were calculated by combining foldChange_score and ppi_score in a new “enhanced_score.” Co-expressed gene clusters were identified using MCL clustering (ENRIGHT; VAN DONGEN; OUZOUNIS, 2002; VAN DONGEN; ABREU-GOODGER, 2012), applied to the attribute-enhanced network, with inflation

hyper-parameter tuned to maximize modularity score (Q). For each of the identified clusters, we also extracted genes that could serve as “local hubs” based on four different metrics: degree, betweenness, eigenvector and closeness.

For the association between genes and phenotypical data we used BNFinder (FROLOVA; WILCZYŃSKI, 2018), combining per-sample normalized gene expression (from RNASeq) with physiological data (derived from HPLC), and converted the observations into classes with the Sturges’ rule (STURGES, 1926).

A two-fold strategy was applied to generate cluster functional labels: the first was based on gene ontology enrichment classes, with the most significant enriched class (lowest FDR) that represented at least 50% of genes within the cluster; the second strategy was based on a Bayesian inference of association with physiological data, with genes being able to be associated either positively or negatively with the changes in each measured metabolite.

Short variant discovery

Short variants (SNPs and Indels) were identified using GATK4 pipeline, in accordance to the best practices for RNASeq short variant discovery (GATK, 2020; VAN DER AUWERA et al., 2013). Aligned RNA/seq reads in BAM format were used as input, as well as GFF and VCF files for R64-1-1 annotation obtained from Ensembl (see “RNASeq alignment and genome annotation”.) Short variant impact was estimated using Ensembl Variant Effect Predictor (MCLAREN et al., 2016).

Multi-omics network assembly

A graph-based approach was used to integrate all data layers (gene expression, co-expressed cluster hubs, pathway impact and nucleotide variants) into a unified network model using NetworkX (HAGBERG; SCHULT; SWART, 2008). The information for each layer was re-structured and merged with the others in way that for every pair of vertices \mathbf{u} and \mathbf{v} , the first vertex (\mathbf{u}) represents a gene and the second vertex (\mathbf{v}) represents the characteristic associated to that gene (a pathway, phenotype or mutation). The weight of the edge defined by (\mathbf{u}, \mathbf{v}) was assigned according to the relationship between the expression change of the gene and alteration on the

pathway/phenotype: +1 for direct relationships (the direction of the gene fold change is in the same direction of the altered pathway/phenotype, i.e. both upregulated), -1 for inverse relationships (the direction of the gene fold change is in the opposed direction of the altered pathway/phenotype, i.e. one upregulated while the other is downregulated) or 0 for neutral relationships (in the case of nucleotide variants). Directionality of the network was established in accordance with the following structure: variant → gene → pathway/phenotype, to reflect the idea that: “a nucleotide variant may affect the gene function, leading downstream alterations”.

Comparative genomics of SA-1 genes

In order to perform comparative genomics analysis for the SA-1 strain with other brazilian bioethanol strains, we developed a web service named INDYdb database (CIAMPONI et al., 2022). Using the integrated API, we performed a query to extract all information for annotated genes of bioethanol-related strains (application=ethanol). An in-depth methodology and usability of this tool will be presented in the second chapter of this thesis (Chapter 2 - INDYdb: An integrative comparative genomics database for industrial yeast strains). The resulting JSON file was then loaded as a dataframe into a python notebook environment (KLUYVER et al., 2016; VAN ROSSUM G; DRAKE FL., 2019, p. 3) for further processing.

Out of the 37 queryable fields in the database, we selected 14 which could be used to quantify differences between SA-1 vs. 6 other brazilian strains (BG-1, CAT-1, JAY291, PE-2.H3, PE-2.H4 and VR-1) and 5 non-brazilian strains (EthanolRed, FaliES1, NCIM3186, ThermosaccDry, ZTW1). These features could be classified in two categories, 6 are boolean values (“perfect match to the reference sequence”, “valid functional annotation”, “presence of a valid ORF”, “absence of a start codon”, “absence of a stop codon” and “presence of an in-frame stop codon”) and 8 are numerical values (“percentage of similarity with reference sequence”, “percentage of exon coverage”, “number of copies”, “number of ORFs”, “number of variant sites”, “gene length”, “transcript length” and “protein length”). Additionally, we used INDYdb data repository to obtain VCF files for each of the 6446 annotated genes and extracted all variant sites identified in SA-1 and the other strains and converted the presence/absence of variation into boolean values (1 and 0). In total 450805 genomic characteristics,

henceforth referred to as features, (14 features for each of the 6446 genes and the variant sites) were teste using a one-sample Student's T-Test (STUDENT, 1908) to compare the SA-1 strain with the other brazilian strains and non-brazilian strains.

RESULTS

Metabolic changes in SA1 yeasts upon pCA exposure

A thorough understanding of the effects of pCA on yeast metabolism is required to generate potential metabolic engineering strategies that can improve strain robustness. Since lignocellulosic hydrolysates contain a high number of phenolic compounds of which 80% represents pCA, this compost was added to the feed-medium of carbon-limited chemostats. A cultivation without inhibitors served as control. Data collected along each batch phase was linearized applying the natural logarithm to exit of CO₂ values as a function of time. Specific consumption rates of glucose and specific production rates of selected extracellular metabolites are shown in Table 1 and Table S1.

Conditions / Parameters	Control	7 mM pCA
μ (batch phase)	0.38 ± 0.02	0.37 ± 0.03
Residual glucose (g / L)	0.69 ± 0.05	0.69 ± 0.22
q glucose	-5.80 ± 0.05	-7.34 ± 0.50
q CO ₂	9.82 ± 1.01	11.04 ± 1.03
q ethanol	8.66 ± 0.34	13.27 ± 1.17
q glycerol	1.04 ± 0.07	0.84 ± 0.24
q lactate	0.08 ± 0.00	0.06 ± 0.03
q pyruvate	0.02 ± 0.00	0.03 ± 0.02
q acetate	0.00 ± 0.00	0.000 ± 0.000
X	2.645 ± 0.007	2.145 ± 0.02
$Y_{X/S}$	0.13 ± 0.00	0.09 ± 0.00
$Y_{Eth/S}$	0.38 ± 0.02	0.46 ± 0.01
C recovery	98.51 ± 3.35	100.761 ± 0.01

Table 1. Physiology of *S. cerevisiae* strains in glucose-limited anaerobic chemostats at a dilution rate of 0.1 h⁻¹. Specific rates (q) are given in mmol g⁻¹ h⁻¹, μ in h⁻¹, X in g DW L⁻¹, $Y_{X/S}$ in g DW g glucose⁻¹

¹, $Y_{Eth/S}$ in g ethanol⁻¹ g glucose⁻¹, and C recovery in (%). The μ value represents the growth rate measured during the batch phase before the steady-state is achieved. Data is the average values of duplicate experiments \pm deviation of the mean.

Differences in the physiological parameters were resulting from the addition of the phenolic compost. Some specific consumption and production rates increased, such as for glucose (26%), CO₂ (12%) and ethanol (53%). On the other hand, we observed a decrease in biomass yield (22%), and in the glycerol production rate (19%) (Table 1). In anaerobic glucose-limited chemostat cultures of the *S. cerevisiae* strains, carbon is mainly diverted to ethanol and CO₂, and minor amounts of glycerol, lactic and acetic acids, with a concomitant formation of yeast biomass. The ethanol yield of SA-1 in the control condition was 21% lower than in the presence of pCA. We found that under anaerobic glucose-limited chemostat cultivations pCA is not metabolized by SA-1 (Figure 1).

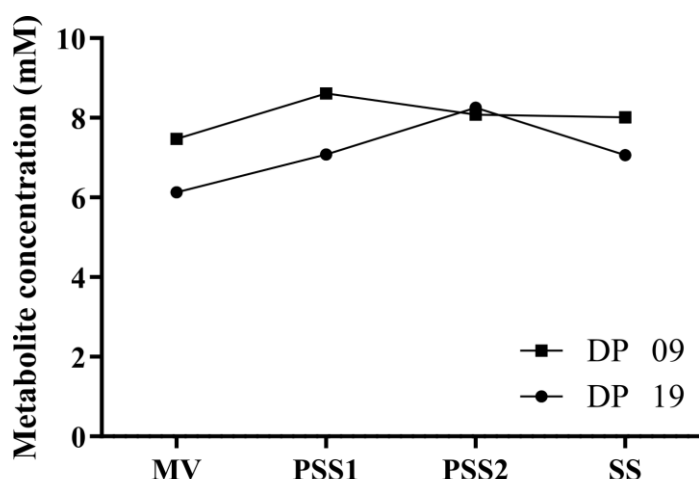


Figure 1. pCA concentration during anaerobic glucose limited chemostat cultivations of *S. cerevisiae* SA-1 strain. MV, medium vessel; PSS1, first pre-steady-state; PSS2, second pre-steady state; SS, steady-state. DP09 and DP19 code for the duplicate runs. The pre-steady-state samples (PSS1 and PSS2) represent samplings at 24 and 48 h after starting the feeding, respectively.

Differential gene expression between treated and control samples

Gene expression analysis based on RNASeq data revealed that both conditions (treated and control) have high correlation between biological duplicates (Pearson's $R^2 > 0.99$, Figure 2A). These values are within the established parameters for chemostat cultures (NOOKAEW et al., 2012), which generate replicates with low biological variability. Additionally, the principal component analysis showed that 79.57% of explained variance observed in the samples can be associated with the axis that represents separation based on experimental conditions, with only 10.28% of

explained variance being associated with alterations within samples under the same conditions (Figure 2B). These analyses indicate significant changes in the transcriptomic landscape of SA1 yeasts when exposed to pCA. We identified a total of 1472 differentially expressed genes between conditions (404 up- and 1068 down-regulated); however, only 448 (10 up- and 438 down-regulated) of the genes had $|\log_2(\text{FoldChange})| \geq 1$ (Figure 3A, Table S2).

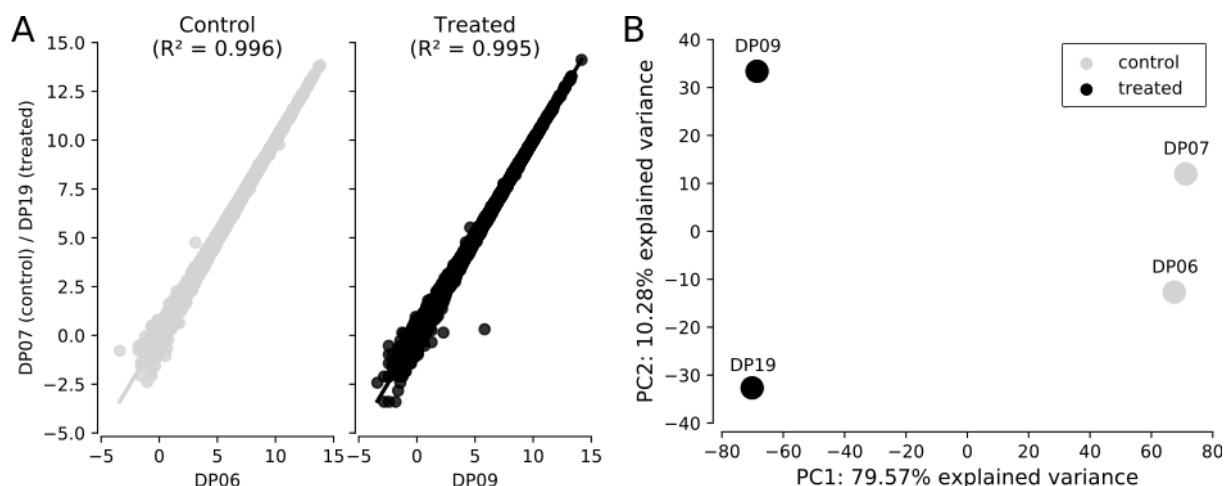


Figure 2. SA1 yeasts show increased production of ethanol and lower biomass yield followed by alterations in the transcriptome. (A) Regression plot showing the relationship (in Pearson's R^2) between biological replicates under the same condition. Each point represents the expression of a gene in $\log_2(\text{CPM})$, measured in the first biological replicate (X-axis) and in the second replicate (Y-axis). **(B)** PCA plot showing the explained variance in the first two components based on normalized gene expression for control (grey) and treated (black) samples.

Our results were mapped in the KEGG database (KANEHISA et al., 2017) to identify perturbations in pathways containing SA-1 differentially expressed gene sets (either up- or down-regulated), which affect the overall pathway activity (Figure 3B, Table S3). Using this approach, we identified a total of 11 pathways that had statistically significant gene set alterations by pCA stress, and three were negatively altered (repressed): oxidative phosphorylation, citrate cycle and peroxisome. The other eight had perturbations with positive effects on the pathway (activated): cell cycle, ribosome biogenesis, metabolic pathways, biosynthesis of amino acids, pyrimidine metabolism, purine metabolism, methane metabolism and glycine, serine and threonine metabolism. Additionally, we identified a total of 20 differentially expressed genes (17 downregulated and 3 upregulated) associated with redox regulation and/or

response to reactive-oxygen species and another set of 15 DEGs (13 downregulated and 2 upregulated) that can be linked to ethanol metabolism and/or fermentation (Figure 3C). We did not observe overlapping genes between these two sets (ROS/Redox and Ethanol/Fermentation).

We then applied graph network clustering using known protein-protein interaction data associated with fold changes derived from RNASeq to identify co-expressed gene clusters. We found a total of 98 clusters within the differentially expressed genes, with 50% of DEGs being located in the 7 biggest clusters (Figure 3D, Table S4). These clusters were then characterized according to their expression profile (Table S5), based on the distribution of $\log_2(\text{foldchange})$ for genes within each cluster, enrichment of significant Gene Ontology classes and KEGG pathways (Table S5). Additionally, no correlation ($R^2 = 0.033$) was found between the number of genes and the overall standard deviation of foldchanges observed within each cluster.

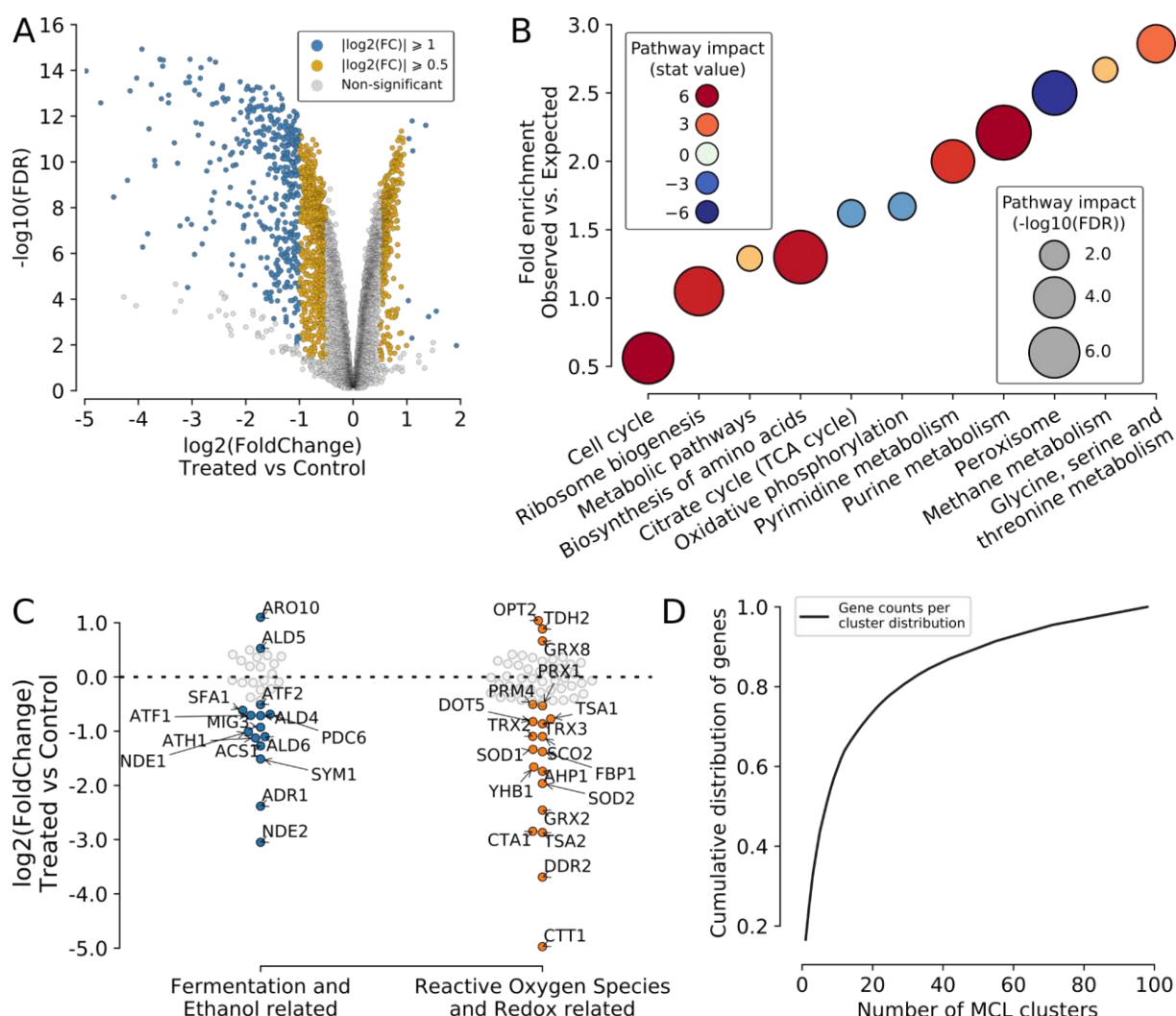


Figure 3. Differentially expressed genes under p-Coumaric stress are skewed towards downregulation and can be grouped into functional clusters. (A) Volcano plot showing the relation between $\log_2(\text{Fold Change})$ (X-axis) and $-\log_{10}(\text{FDR})$ (Y-axis) for differentially expressed genes (DEGs in blue and gold) after pCA treatment. **(B)** Point plot showing the predicted perturbations in KEGG pathways (X-axis) based on RNASeq data. The Y-axis shows the fold-enrichment score identified for each pathway of DEGs; the size of the point shows the significance value of the prediction (in $-\log_{10}(\text{FDR})$ scale), and the color represents the direction of perturbation: downregulated (brown) or upregulated (blue). **(C)** Swarmplot showing the $\log_2(\text{Fold Change})$ profile (Y-axis) of the 13 DEGs associated with fermentation and/or ethanol (blue) and the 17 DEGs associated with reactive oxygen species response and redox processes (orange). **(D)** Cumulative distribution function (CDF) plot showing the percentage of DEGs (Y-axis) that are in clusters (X-axis), starting from the biggest clusters (in number of genes).

Functional characterization of co-expressed gene clusters

In order to further explore these gene sets, we filtered the clusters in which the expression values located in a distance of $1.5 * \text{IQR}$ (interquartile range) were in the same quadrant, either above or below zero, and had more than 20 genes. A total of 9 clusters (C2, C3, C4, C6, C7, C9, C10, C11 and C12), with a total of 462 DEGs, were selected and characterized according to biological functional enrichment and association (Figure 4A, Table S6) with phenotypical alterations (Figure 4B, Table S7). Six of these clusters (C2, C4, C7, C9, C10 and C11) comprised downregulated genes, while 3 (C3, C6 and C12) were mostly from upregulated genes (Figure 4C). However, each cluster had distinct associations with biomass yield and ethanol production (Figure 4D). Downregulated clusters tended to share positive associations with biomass and negative relations with ethanol production. Upregulated clusters, on the other hand, showed an inverse pattern, with negative regulation of biomass and positive association with ethanol. To further explore the characterized clusters, we also evaluated each of the genes found within the clusters to establish which of them acted as “network hubs” in their respective clusters, based on their values for eigenvector centrality, betweenness, degree and closeness. In total, we identified 25 genes (Figure 5, Table 2) that act as main points of interaction for the genes in the network.

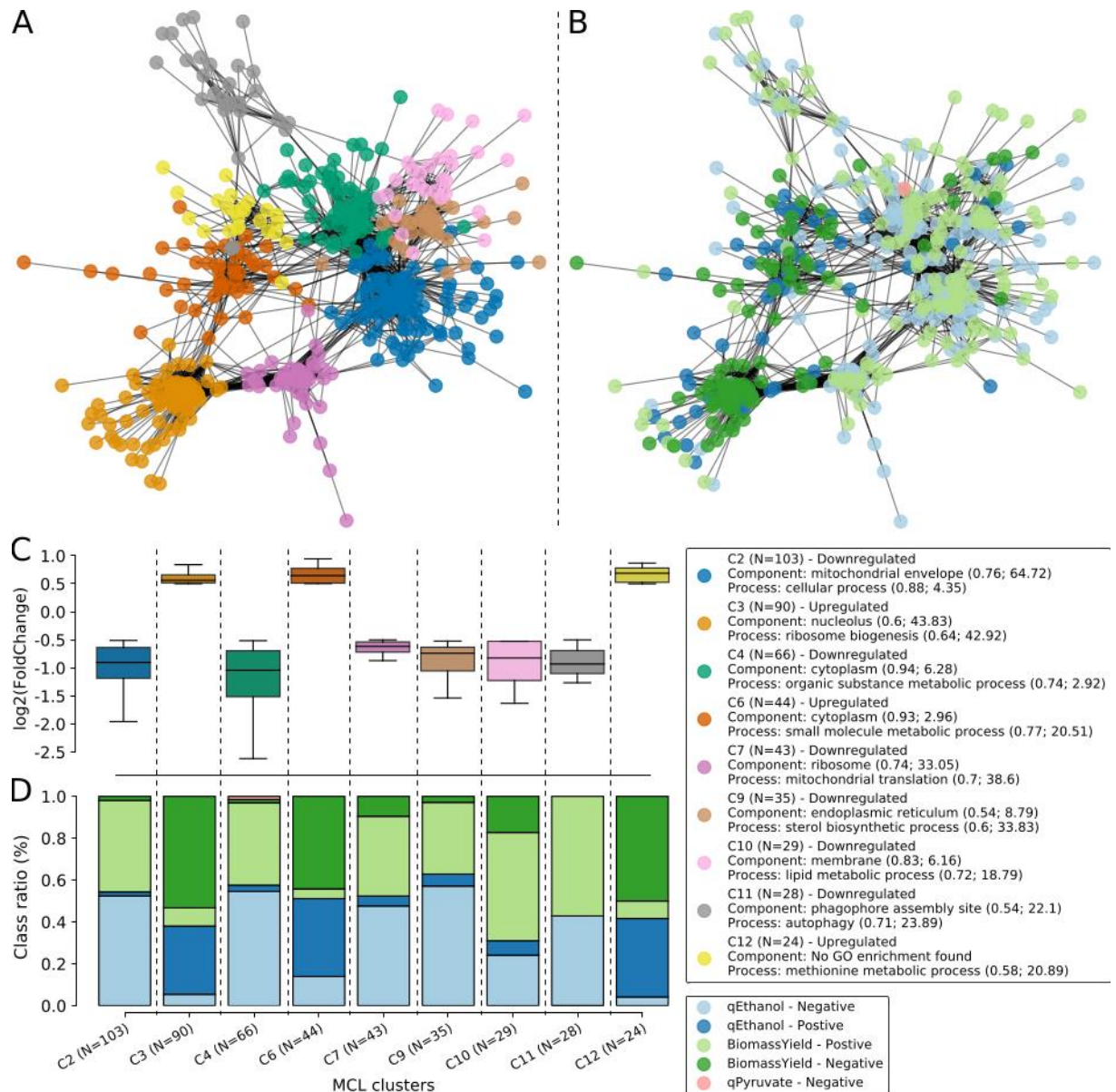


Figure 4. Co-expressed gene clusters can be classified according to functional enrichment and association with phenotype alterations. (A-B) Network representation of the 9 co-expressed gene clusters selected for further exploration. On the right (A), the clusters were classified according to their gene ontology functional enrichment; on the left (B) each gene in the network was classified according to their probability of having a positive or negative association with either qEthanol or biomass yield metabolic phenotypes. (C) Boxplot showing the expression profile (Y-axis) of each of the selected clusters (X-axis). The color of each cluster matches those of the network. (D) Stacked barplot showing the ratio of genes (Y-axis) associated with the metabolic classes (positive/negative relation to biomass yield/qEthanol) for each of the selected clusters.

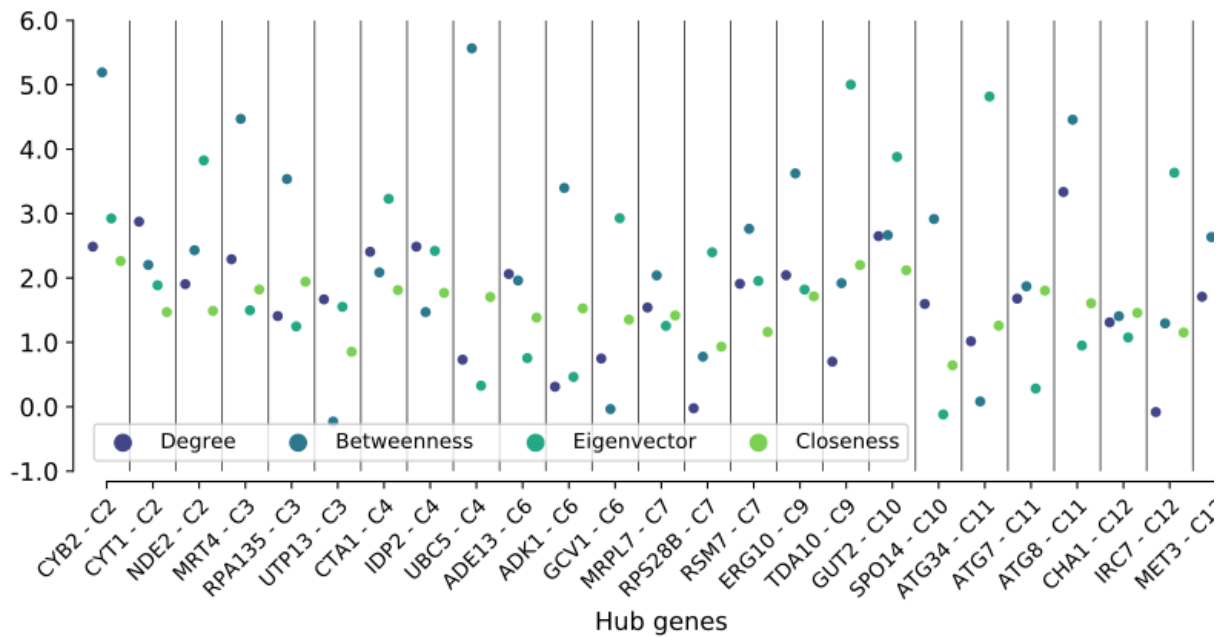


Figure 5. Hub genes located in co-expressed gene clusters. Swarmplot showing the centrality scores (Y-axis) measured for the hub-genes selected from each cluster (X-axis). Each metric used is shown in a different color.

Gene	Cluster	Description (UniProt Keywords)
CYB2	C2	Electron transport, FMN, Flavoprotein, Heme, Iron, Metal-binding, Mitochondrion, Oxidoreductase, Respiratory chain, Transit peptide, Transport
CYT1	C2	Electron transport, Heme, Iron, Membrane, Metal-binding, Mitochondrion, Mitochondrion inner membrane, Respiratory chain, Transit peptide, Translocase, Transmembrane, Transmembrane helix, Transport
NDE2	C2	FAD, Flavoprotein, Mitochondrion, NAD, Oxidoreductase, Transit peptide
MRT4	C3	Cytoplasm, Nucleus, Ribosome biogenesis
RPA135	C3	Acetylation, DNA-directed RNA polymerase, Metal-binding, Nucleotidyltransferase, Nucleus, Phosphoprotein, Ribosome biogenesis, Transcription, Transferase, Zinc, Zinc-finger
UTP13	C3	Nucleus, Repeat, Ribonucleoprotein, Ribosome biogenesis, WD repeat, rRNA processing
CTA1	C4	Acetylation, Heme, Hydrogen peroxide, Iron, Metal-binding, Oxidoreductase, Peroxidase, Peroxisome
IDP2	C4	Cytoplasm, Glyoxylate bypass, Magnesium, Manganese, Metal-binding, NADP, Oxidoreductase, Tricarboxylic acid cycle
UBC5	C4	ATP-binding, Isopeptide bond, Nucleotide-binding, Phosphoprotein, Stress response, Transferase, Ubl conjugation, Ubl conjugation pathway

ADE13	C6	Isopeptide bond, Lyase, Purine biosynthesis, Ubl conjugation
ADK1	C6	ATP-binding, Acetylation, Cytoplasm, Kinase, Mitochondrion, Nucleotide-binding, Transferase
GCV1	C6	Aminotransferase, Mitochondrion, Transferase, Transit peptide
MRPL7	C7	Mitochondrion, Ribonucleoprotein, Ribosomal protein, Transit peptide
RPS28B	C7	Acetylation, Cytoplasm, Ribonucleoprotein, Ribosomal protein
RSM7	C7	Mitochondrion, Ribonucleoprotein, Ribosomal protein, Transit peptide
ERG10	C9	Acetylation, Acyltransferase, Cytoplasm, Metal-binding, Potassium, Transferase
TDA10	C9	ATP-binding, Cytoplasm, Kinase, Nucleotide-binding, Nucleus, Transferase
GUT2	C10	FAD, Flavoprotein, Membrane, Mitochondrion, Mitochondrion inner membrane, Oxidoreductase, Transit peptide
SPO14	C10	Acetylation, Hydrolase, Lipid degradation, Lipid metabolism, Meiosis, Phosphoprotein, Repeat, Sporulation
ATG34	C11	Autophagy, Membrane, Protein transport, Transport
ATG7	C11	Autophagy, Cytoplasm, Protein transport, Transport, Ubl conjugation pathway
ATG8	C11	Autophagy, Cytoplasmic vesicle, Lipoprotein, Membrane, Protein transport, Transport, Ubl conjugation pathway, Vacuole
CHA1	C12	Acetylation, Lyase, Mitochondrion, Pyridoxal phosphate
IRC7	C12	Amino-acid biosynthesis, Lyase, Methionine biosynthesis, Pyridoxal phosphate
MET3	C12	ATP-binding, Amino-acid biosynthesis, Cysteine biosynthesis, Cytoplasm, Methionine biosynthesis, Nucleotide-binding, Nucleotidyltransferase, Transferase

Table 2. List of genes that were identified as hubs in the co-expressed clusters. Genes identified as interaction hubs in the clusters (Cluster) were annotated using keywords assigned by the UniProt database that reflect their functional and structural characteristics.

Prediction of genomic short-variants based on RNA-seq

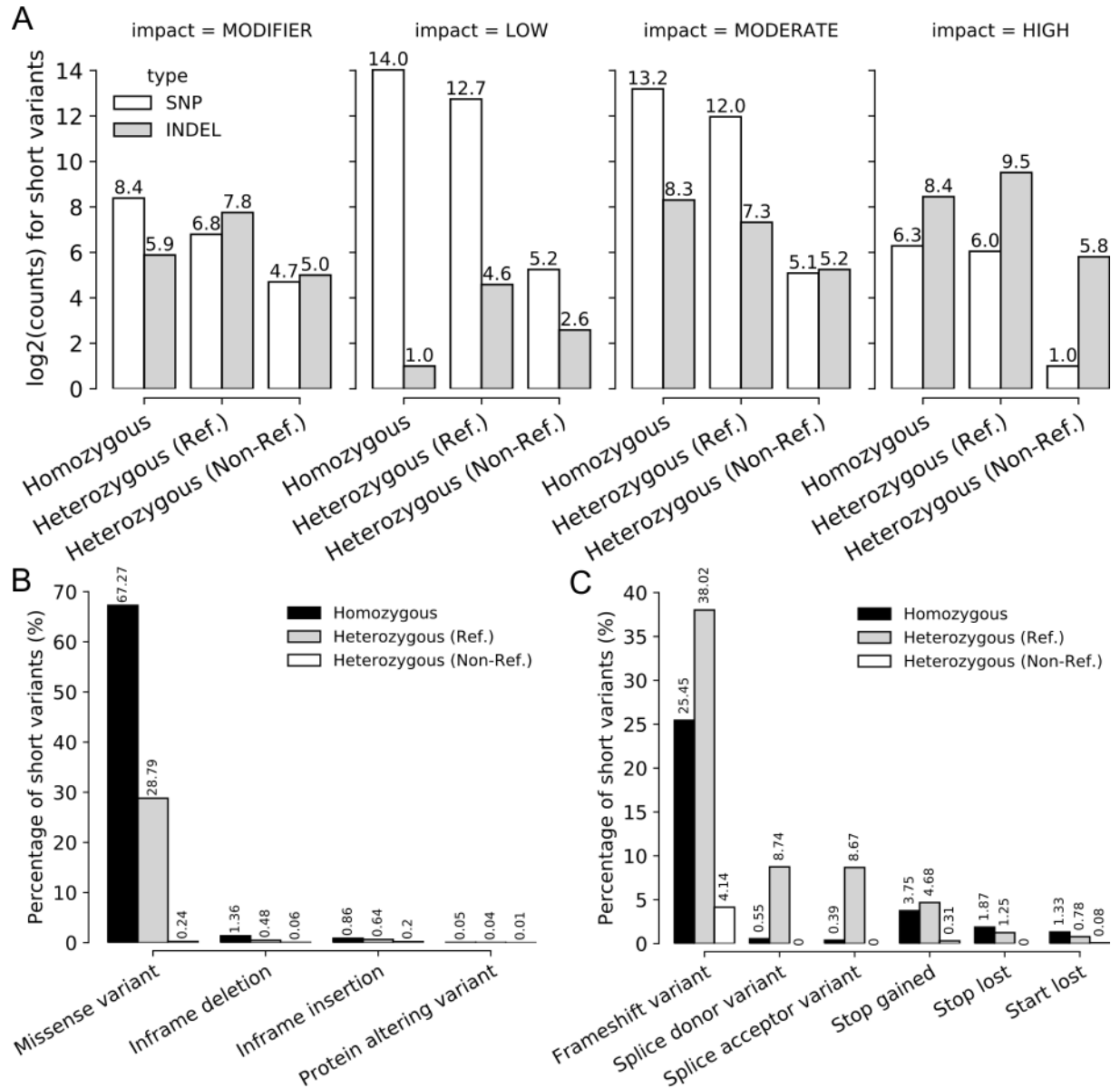
Using data collected from high-throughput RNA sequencing, we reconstructed short variants (SNPs and INDELs) that occur within the transcripts and predicted the impact that they might have on their associated coding sequence (GATK, 2020). We identified a total of 38420 short variants (compared to the R64-1-1 reference

annotation) that can be subdivided into three major categories: homozygous (both alleles carry the variant), heterozygous with reference (one allele carries the variant, while the other is equal to the reference) and heterozygous without reference (both alleles carry different variants, and neither is equal to the reference). These short variants were then classified according to their predicted impact on their associated coding sequence (Figure 6A): 683 modifier variants (non-coding, e.g., intronic or UTR variants); 23388 low impact variants (e.g., synonymous mutations); 13655 moderate impact variants (e.g., missense mutations that preserve overall protein length/structure) and 1093 high impact variants (e.g., frameshift INDELs or stop-gaining SNPs). However, the sum of the subclasses exceeds the total of variants (Table S8). This occurs because a variant can impact multiple genes, and current limitations make it difficult to solve such conflicts using HTS data alone (GATK, 2020; VAN DER AUWERA et al., 2013).

The major classes of high impact variants involved frameshift variants, and most of them (~38%) were heterozygous (in relation to the reference) in nature. Homozygous frameshift variants also comprised the second biggest class, with approximately 25% of variants falling in that class, and non-reference heterozygous frameshifts comprising ~4% of high impact variants. Heterozygous variants located in splicing sites were also a major class of high impact predictions, with a combined total of ~17%. Lastly, stop-gaining mutations represented ~8% of all high impact variants, with approximately half (3.75%) being homozygous in nature and the remaining (4.68%) being heterozygous.

We then filtered only the variants that had predicted moderate and high protein impacts for further exploration. Our analysis showed that the major class of moderate impact alterations was the class of missense variants, with ~67% of them being homozygous in nature and ~29% being heterozygous with the reference (Figure 6B). However, when we compared the same results for high impact variants, we observed a much broader distribution of CDS consequences (χ^2 (3, N = 15166) = 7096.5, $p < .001$, Figure 6C).

Figure 6. Homozygous missense and heterozygous frameshift variants are the major classes of short-variants with predicted moderate-to-high protein impact. (A) Barplots showing the overall



Moderate predicted protein impact
number of variants (Y-axis) identified from RNASeq data for the SA1 strain when compared to the R64-1-1 reference annotation for *S. cerevisiae*. The X-axis shows the ploidy identified for each variant: homozygous, heterozygous with reference, or heterozygous without reference. The columns are separated according to overall level of predicted protein impact, from lowest to highest. **(B-C)** Barplot showing the breakdown of the predicted protein impacts (X-axis) and their associated proportions (Y-axis) for variants with moderate **(B)** and high **(C)** predicted protein impact.

Assembly of a multi-omics network model for pCA response

In order to generate a single model that represents the association between differentially expressed gene located in perturbed pathways, the association with fermentation/ethanol and ROS/Redox, presence of high-impact short variants and phenotype impact prediction, we converted all the information described in the previous sections into a graph-based network format. This created a comprehensive panorama of the interactions occurring during pCA stress response (Table S9). From this network, we extracted all the edges in which at least one vertex was either a hub gene (as shown in Figure 5) or a gene associated with fermentation/ethanol or ROS/Redox (as shown in Figure 3C). A total of 16 genes (Figure 7, Table 3) were selected based on the aforementioned criteria for constructing the model, while these targets were clustered into two major groups: those associated with ethanol production (IDP2, ERG10, CYT1, ARO10, GCV1, TDA10 and CHA1) and those related to biomass yield (SOD1, CTA1, IRC7, SPO14, UTP13, CYB2, MET3, ADK1 and ADE13).

This multi-level network also showed the type of interaction between genes and their targeted phenotype and associated pathway, which can be either a direct relationship between the changes in gene expression and the pathway/phenotype alteration (e.g. both upregulated) or an inverse relationship (where one is upregulated and the other is downregulated). In total, we found 19 positive interactions (where the gene foldChange occurs in the same direction as the change in pathway activity or metabolite measurement) and 11 negative interactions (where the gene foldChange occurs in the opposite direction as the change in pathway activity or metabolite measurement). Our network showed 6 genes with predicted mutations (TDA10, CHA1, SPO14, IRC7, ADK1 and ADE13), 2 genes associated with ROS/Redox processes (CTA1 and SOD1) and 1 gene directly related to fermentation/ethanol (ARO10).

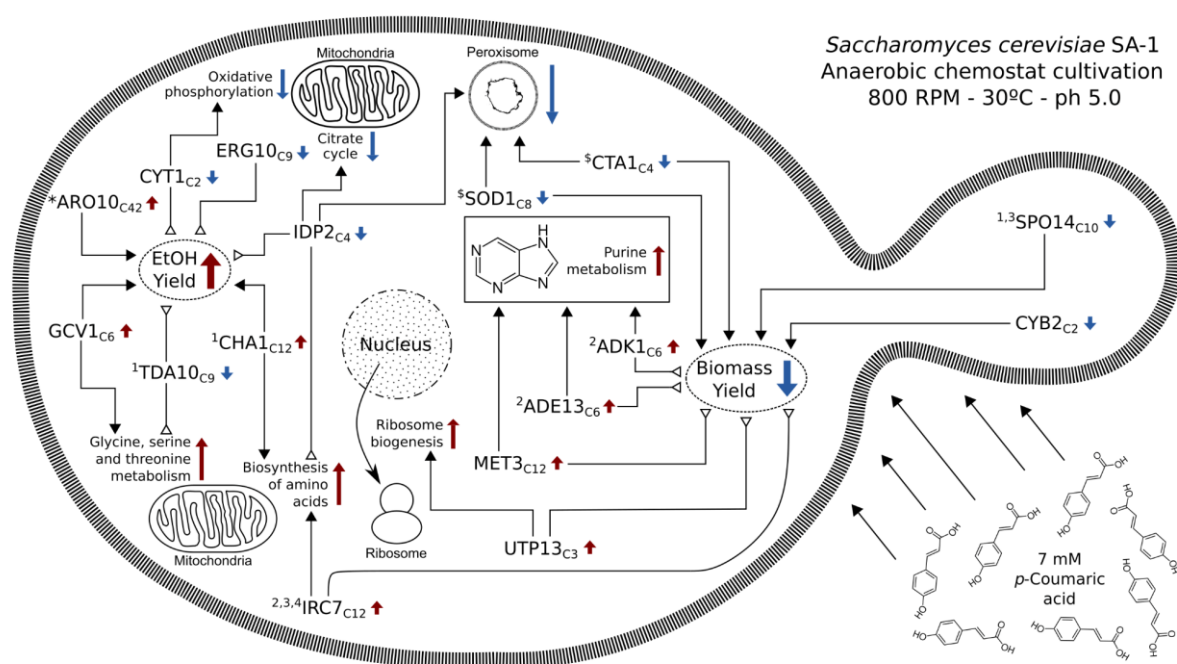


Figure 7. Multi-omics integrated model for alterations induced by pCA stress in the *S. cerevisiae* SA1 strain. Network representation of the relation between hub genes, short variants (SVs), perturbed pathways and phenotypic alterations. Red arrows indicate positive regulation (upregulated genes and pathways), blue arrows show negative relations (downregulated genes and pathways) and black arrows represent additional information sources (presence of SVs or gene ontology categories), with their connector showing the relationship as direct (solid arrowhead, both elements moving in the same direction) or inverted (hollow triangle, elements moving in opposite directions). Genes labeled with numbers indicate the presence of a short variant (1: missense homozygous mutation; 2: missense heterozygous; 3: frameshift heterozygous and 4: stop-gained heterozygous). Genes labeled with symbols represent known associations to either fermentation processes (*) or response to reactive oxygen species (\$).

Gene	Cluster	Phenotype	Pathway	SNV	Description (UniProt Keyword)
UTP13	C3	Biomass (-)	Ribosome biogenesis (+)	No	Nucleus, Repeat, Ribonucleoprotein, Ribosome biogenesis, WD repeat, rRNA processing
ADE13	C6	Biomass (-)	Purine metabolism (+)	Yes	Isopeptide bond, Lyase, Purine biosynthesis, Ubl conjugation
ADK1	C6	Biomass (-)	Purine metabolism (+)	Yes	ATP-binding, Acetylation, Cytoplasm, Kinase, Mitochondrion, Nucleotide-binding, Transferase

IRC7	C12	Biomass (-)	Biosynthesis of aminoacids (+)	Yes	Amino-acid biosynthesis, Lyase, Methionine biosynthesis, Pyridoxal phosphate
MET3	C12	Biomass (-)	Purine metabolism (+)	No	ATP-binding, Amino-acid biosynthesis, Cysteine biosynthesis, Cytoplasm, Methionine biosynthesis, Nucleotide-binding, Nucleotidyltransferase, Transferase
CYB2	C2	Biomass (+)	None	No	Electron transport, FMN, Flavoprotein, Heme, Iron, Metal-binding, Mitochondrion, Oxidoreductase, Respiratory chain, Transit peptide, Transport
CTA1	C4	Biomass (+)	Peroxisome (+)	No	Acetylation, Heme, Hydrogen peroxide, Iron, Metal-binding, Oxidoreductase, Peroxidase, Peroxisome
SOD1	C8	Biomass (+)	Peroxisome (+)	No	Antioxidant, Copper, Cytoplasm, Disulfide bond, Isopeptide bond, Metal-binding, Mitochondrion, Oxidoreductase, Phosphoprotein, Ubl conjugation, Zinc
SPO14	C10	Biomass (+)	None	Yes	Acetylation, Hydrolase, Lipid degradation, Lipid metabolism, Meiosis, Phosphoprotein, Repeat, Sporulation
CYT1	C2	Ethanol (-)	Oxidative phosphorylation (+)	No	Electron transport, Heme, Iron, Membrane, Metal-binding, Mitochondrion, Mitochondrion inner membrane, Respiratory chain, Transit peptide, Translocase, Transmembrane, Transmembrane helix, Transport
IDP2	C4	Ethanol (-)	Peroxisome (+), Citrate cycle (+), Biosynthesis of aminoacids (-)	No	Cytoplasm, Glyoxylate bypass, Magnesium, Manganese, Metal-binding, NADP, Oxidoreductase, Tricarboxylic acid cycle
ERG10	C9	Ethanol (-)	None	No	Acetylation, Acyltransferase,

					Cytoplasm, Metal-binding, Potassium, Transferase
TDA10	C9	Ethanol (-)	Glycine, serine and threonine metabolism (-)	Yes	ATP-binding, Cytoplasm, Kinase, Nucleotide-binding, Nucleus, Transferase
GCV1	C6	Ethanol (+)	Glycine, serine and threonine metabolism (+)	No	Aminotransferase, Mitochondrion, Transferase, Transit peptide
CHA1	C12	Ethanol (+)	Biosynthesis of aminoacids (+)	yes	Acetylation, Lyase, Mitochondrion, Pyridoxal phosphate
ARO10	C42	Ethanol (+)	None	No	Branched-chain amino acid catabolism, Cytoplasm, Decarboxylase, Isopeptide bond, Lyase, Magnesium, Metal-binding, Phenylalanine catabolism, Thiamine pyrophosphate, Tryptophan catabolism, Tyrosine catabolism, Ubl conjugation

Table 3. List of genes that were used as nodes in the multi-omics model. Genes identified as anchor nodes in the multi-omics model were annotated using keywords assigned by the UniProt database that reflect their functional and structural characteristics. The additional columns also show their predicted phenotypic association (Phenotype), any associated pathways (Pathway) and if the gene contains a SNV site (SNV).

Identification of distinguishable genomic features for SA-1 yeast strain

In order to further explore the hub gene network constructed for p-Coumaric stress response, we compared the genomic features found in protein-coding genes of the SA-1 strain with the same genes found in 6 other bioethanol-related brazilian strain (CAT-1, VR-1, BG-1, JAY291, PE-2.H3 and PE-2.H4) and 5 non-brazilian strains (Fali ES1, Thermosacc Dry, NCIM3186, Ethanol Red and ZTW1).

A total of 450805 genomic features were extracted from INDYdb (CIAMPONI et al., 2022), ranging from structural gene characteristics (such as gene/transcript/protein length, presence and validity of ORFs, number of variants) to sequence-specific features (percentage identity with S288C reference, percentage of exon coverage, number of nucleotide variant sites, etc...). Additionally, we also

obtained data regarding mutations associated with changes in the protein sequence by extracting the data contained in the VCF files available in the data repository that were derived from multiple-sequence-alignment of predicted protein sequences for the genes annotated in the database. This step was performed using a GET request via the integrated API that performed an SQL query on INDYdb's database and obtained the information for the genomic features of interest for every gene of every strain associated with bioethanol production. The numerical values representing each feature were split into three distinct groups: SA-1 strain, other brazilian strains and non-brazilian strains. We then applied an iterative approach that compared, for each feature of each gene, the population mean values of the "other brazilian strains" and "non-brazilian" with the value observed for the SA-1 strain.

A total of 7035 features were represented statistically significant alterations between SA-1 and the other brazilian bioethanol-related strains (Figure 8A, Table S10). We then filtered the genes containing these features using data from the co-expressed gene clusters and from the hub genes and their known interactors. In total, we found 67 genes that fitted the three criteria established: Containing a distinguishable feature, present in one of the co-expressed gene clusters and being either a hub gene or a direct interactor of one (Figure 8B). Three hub genes (SPO14 and TDA10 and GCV1) were found in that group, while the remaining 64 genes represented interactors of these hubs (in alphabetical order: AAC1, ADE17, ADE4, AFG2, ANT1, ATG2, ATG7, ATP11, CAT8, CEM1, COQ3, COX10, COX11, COX8, DBP7, DHR2, EHD3, ERG26, ERG27, ERG3, ERG7, FOX2, GLY1, GPD2, HFD1, ICL1, IMD2, JEN1, KRE33, MDH1, MDH2, MET17, MRH4, MSS51, NDE1, NDE2, NEW1, NRP1, PAH1, PEX14, POT1, PRP43, PXA1, PXA2, QCR2, REX4, RIO2, RTT10, SER2, SIP4, SQT1, SSF1, TAZ1, THR1, TIM13, TRM44, TSR2, UBC5, URB2, UTP4, UTP8, UTP9, YAT2 and YJU3).

The majority of genes (47) in the intersection group contained variants in the protein sequence, while 42 genes showed an overall difference in the number of accumulated single-nucleotide variant sites. Four genes (PAH1, SIP4, TAZ1 and IMD2) showed differences in overall gene and protein length, with SIP4 and IMD2 showing less coverage of exons (in comparison with S288C strain). Additionally, IMD2,

which is a direct interactor of the ADE13 hub gene, was also predicted to have a non-functional ORF due to a lack of stop codon in its transcript sequence (Figure 8C). Functional analysis of the gene set containing all 67 identified genes revealed an statistically significant enrichment of 12 different KEGG pathways (Figure 8D), with the majority of them (8) being associated with metabolic processes. Individual targets can then be further explored using multiple sequence alignment in order to identify potential candidates (either whole genes or specific target regions) for genetic engineering approaches, here exemplified by the protein sequence variations in SPO14 (hub gene with distinguishing features, Figure 9A), IRC7 (hub gene without distinguishing features, Figure 9B) and ORF changes in IMD2 (Figure 9C).

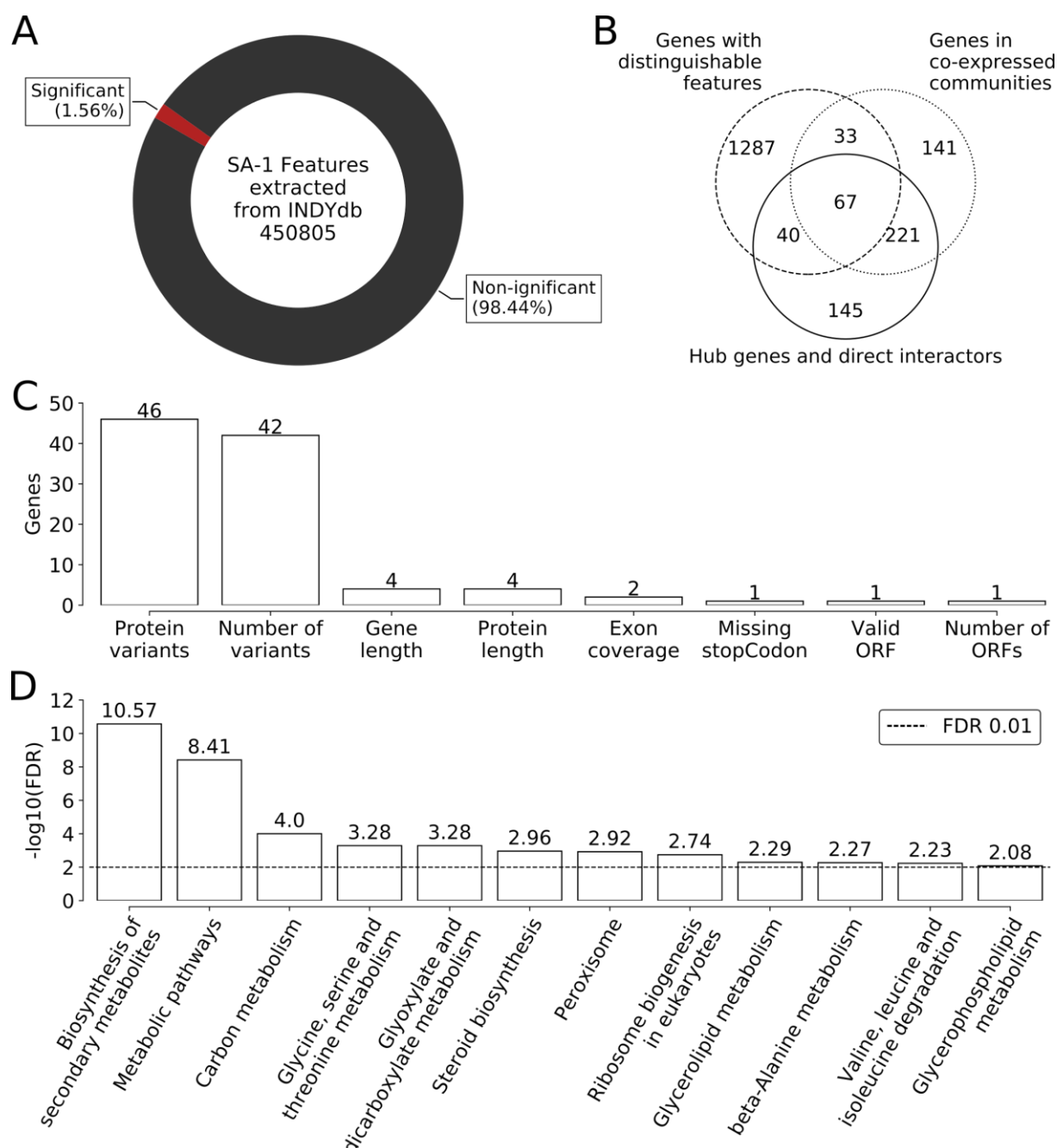
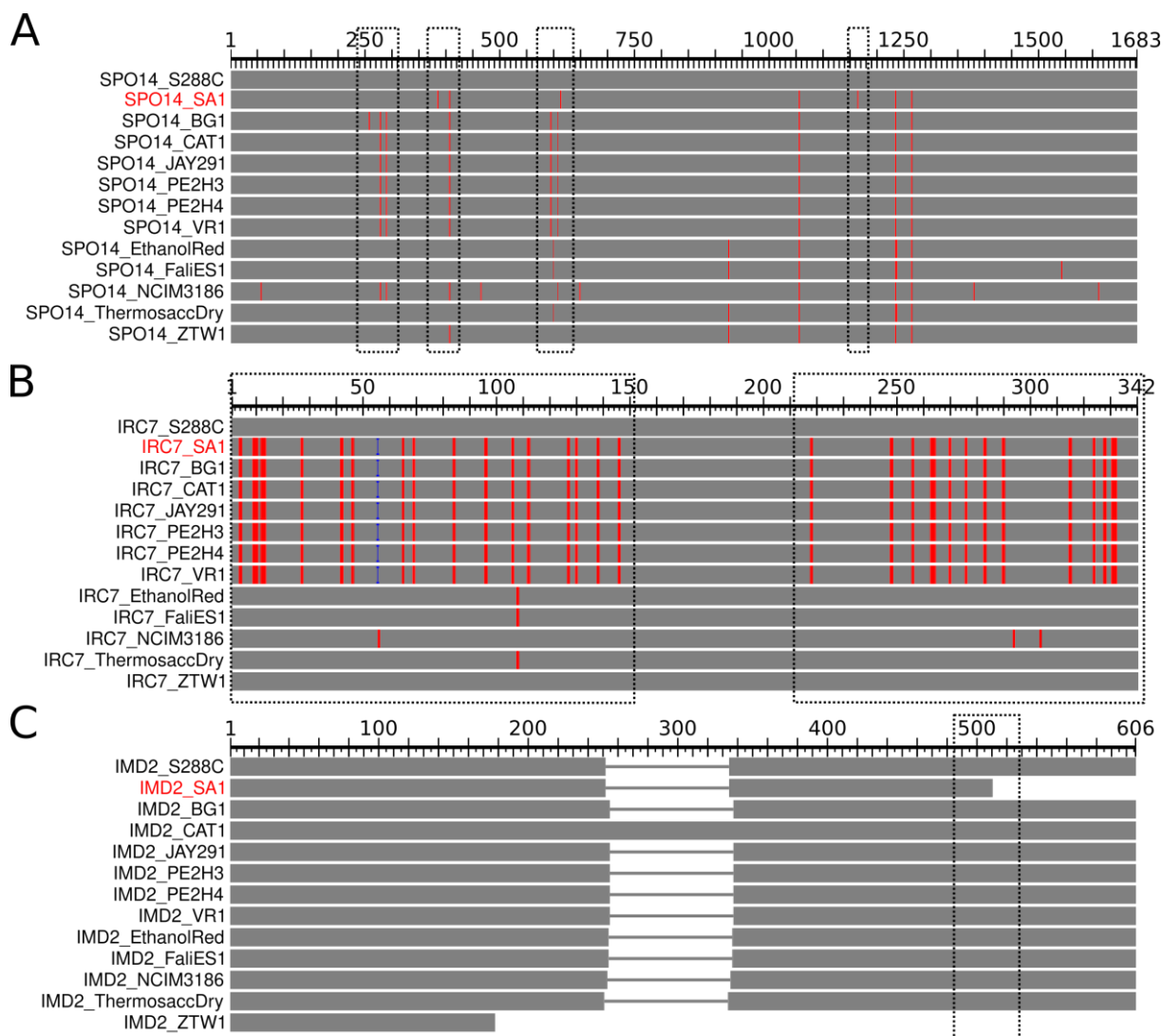


Figure 8. Distinguishing genomic features of genes associated with SA1 strain para-coumaric acid response. (A) Plot showing the ratio between statistically significant (red) and non-significant (grey) features that distinguish SA-1 genomic makeup from the other Brazilian bioethanol-related strains. **(B)** Venn diagram showing the intersection between genes containing distinguishable features (dashed), genes located in co-expressed clusters (dotted) and hub genes and their interactors (solid). **(C)** Barplot showing the number of genes (Y-axis) in each of the classes of genomic features (X-axis) found in the intersection of the three groups described in the Venn diagram. **(D)** Barplot showing the FDR value of the enrichment analysis (Y-axis, in $-\log_{10}$ scale) for the statistically significant KEGG pathways found (X-axis).

Figure 9. Multiple sequence alignment allows direct comparison of SA-1 targets with other bioethanol-related industrial strains. Graphical representation of the multiple sequence alignment of



the protein sequences for three distinct gene products: SPO14 (**A**), IRC7 (**B**) and IMD2 (**C**). The grey bars represent protein sequence (X-axis) across multiple yeast strains (Y-axis), with the scale above showing the overall length of the biggest sequence. The red regions mark the occurrence of variation sites (changes in amino-acids) in comparison with S288C strain and the dotted boxes represent regions that contain genomic features of interest for the SA-1 strain.

DISCUSSION

It has been reported that in aerobic cultures containing pCA, the growth rate is significantly reduced in a dose-dependent manner and inhibiting the efficient bioconversion of lignocellulose biomass by fermentative organisms (ADEBOYE et al., 2015; ADEBOYE; BETTIGA; OLSSON, 2014). Although the *S. cerevisiae* CEN.PK113-7D strain was shown to be capable of slow in situ catabolic conversion of 7 mM pCA in aerobic batch cultivations, performing a complete conversion of pCA into other phenolic compounds over a period of 72 h (ADEBOYE et al., 2015; ADEBOYE; BETTIGA; OLSSON, 2014). Our results indicate that under anaerobic conditions, the same catabolic effect does not occur with the pCA concentrations remaining relatively stable across steady-state measurements. This highlights the importance of analyzing industrial *S. cerevisiae* strains under anaerobic chemostat conditions, especially considering that mitochondrial respiration has been shown to be of significant importance for controlling yeast growth rate processes and resistance to phenolic compounds (FLETCHER; BAETZ, 2020; KITAGAKI; TAKAGI, 2014; MALECKI et al., 2020).

Physiological data collected from the anaerobic chemostat cultures showed that pCA exposure may increase ethanol production in SA1 yeasts ($\log_2(\text{foldChange}) = 0.58$; t-test p-value < 0.05). We also observed that these changes are accompanied by an increased glucose uptake ($\log_2(\text{fC}) = 0.33$, pval = 0.10), which might suggest alterations in the metabolic state. In addition, these changes were followed by a decrease in the overall dry mass ($\log_2(\text{fC}) = -0.27$, pval < 0.05) and biomass yield ($\log_2(\text{fC}) = -0.25$, pval = 0.10). This is consistent with previous studies that showed that *S. cerevisiae*, when exposed to pCA during laboratory-controlled batch fermentation, increases overall metabolism and ethanol production (ADEBOYE et al., 2015). Strains resistant to lignocellulosic inhibitors have lower growth rates when exposed to such compounds (GU; ZHANG; BAO, 2014); however, SA1 yeasts showed no signs of slower growth during the batch phase, with both control and treated samples showing an average μ of 0.370.

It is also known that the presence of insoluble lignocellulosic inhibitors in the medium can promote generalized downregulation of gene expression in another industrial *S. cerevisiae* strain (Moreno et al. 2019). With that in mind, we explored the possible transcriptomic alterations induced by pCA in our chemostat experiments. By using gene expression data obtained from our paired RNA-seq data (i.e. from the same sample, we extracted, at the same time, fractions for metabolite analysis as well as RNA extraction and subsequent high-throughput sequencing), we identified the same tendency towards a downregulation of the transcriptional machinery, but we found no expression similarity between the two sets of differentially expressed genes. Out of all DEGs encountered in both studies, only 92 were shared by both strains (less than 5% of the total number of genes identified), and within this subset no correlation coefficient between fold-changes was found (Person's $R^2 = 0.068$).

In order to further improve our ability to understand the transcriptional machinery involved in the response to pCA, we extracted the gene expression information for the both control and treated samples and then projected that information onto KEGG mappings using the fold-change information to predict what would be the biological impact on those pathways. In general, our data suggests that SA1 cell activity may increase during pCA response. This corroborates previous studies that established that yeasts activate multiple regulatory pathways, modulating cell cycle and metabolic rates to compensate for adverse environmental conditions (MORENO et al., 2019). This effect, when coupled with the anaerobic environment used in the chemostat fermentation experiments, might be associated with the increased ethanol production observed.

Amongst the impacted pathways, purine metabolism was one of the most prominent, this pathway is involved in the formation of adenine and guanine. The former is an essential part of the overall cell metabolism, in the form of ATP/ADP/AMP, by providing energy for cellular processes, and the latter plays a distinct role in cell response to stress conditions in the form of GTP/GDP, a molecule often used in signaling processes during stress response for transcriptional regulation (BRAUER et al., 2008) and glucose signaling via GPCR (FUCHS; MYLONAKIS, 2009). The increase in the metabolism of these compounds could also be related to an increased

rate of metabolic processes during stress response. Furthermore, we also identified several differentially expressed genes that have known associations, via Gene Ontology (ASHBURNER et al., 2000; GENE ONTOLOGY CONSORTIUM, 2021), to reactive oxygen species and ethanol fermentative processes. Both of these processes are known to be intricately related to the mitochondria and peroxisome organelles (AYER; GOURLAY; DAWES, 2014; KITAGAKI; TAKAGI, 2014; SIBIRNY, 2016), and are known hallmarks of yeast response to stress induced by lignocellulosic inhibitors (JAYAKODY et al., 2013; LI et al., 2019; SIBIRNY, 2016; VALL-LLAURA et al., 2019). In addition to the aforementioned processes, we also identified several other upregulated pathways that are linked to mitochondrial activity, such as biosynthesis and metabolism of aminoacids (MALINA; LARSSON; NIELSEN, 2018). These processes intrinsically associated with the TCA cycle, which is downregulated in our dataset, and of paramount importance in maintaining amino-acid homeostasis, which, in turn, is vital for promoting long-term viability in yeasts (BACCOLO et al., 2018).

In addition to several of our results suggest towards alterations surrounding pathways related to the mitochondria in response to *pCA* stress, multiple studies already pointed out the importance of this organelle towards the resistance to this particular compound (MORALES; MENDOZA; COTORAS, 2017; MUKAI et al., 2010; RICHARD; VILJANEN; PENTTILÄ, 2015, p. 1). However, the exact mechanisms by which the *pCA* affects this *S. cerevisiae* organelle under anaerobic conditions is still largely unexplored. When exploring the literature available for other organisms, we find studies conducted in rat liver and human cells showed that *pCA* may cause damage to the mitochondria: by inhibiting the pyruvate transport mechanism (KILLEEN; BOULTON; KNOESEN, 2018; LIU; QIN; LIN, 2017), inducing reactive oxygen species (ROS) damage (LIMA et al., 2006) and mitochondrial membrane depolarization (POSADINO et al., 2013). Another study demonstrated that PAD1, a mitochondrial protein that is downregulated ($\log_2(\text{FC}) = -0.5$) in SA1, is essential for the decarboxylation of phenylacrylic acids (MUKAI et al., 2010). In all the three cases described in the literature, mitochondrial damage ultimately leads to a signaling cascade that starts cell autophagy.

In order to delve deeper into the transcriptomic landscape alterations induced by pCA stress on SA-1 yeasts, we shifted our focus to the co-expressed gene clusters extracted from our network in order to characterize which functional groups of genes were being up/down-regulated and if those clusters could be related to the pathways changes observed previously. Amongst the groups of genes that stood out from the background, we noticed that one of the co-expressed clusters associated with autophagy (C11) was actively repressed in SA1 under p-Coumaric stress. Our data also suggests a potential disruption of the mitochondria, with 70% of the genes located in two downregulated clusters being directly associated with mitochondrial cellular processes (C2) and translation (C7). This correlates with the negative impact observed for KEGG pathways associated with citrate cycle and oxidative phosphorylation. Furthermore, we observed that 20 of the 27 differentially expressed genes associated with the peroxisomal pathway (AGX1, CAT2, CTA1, DCI1, ECI1, FAA2, IDP2, IDP3, PEX1, PEX11, PEX14, PEX2, PEX5, POT1, POX1, PXA1, PXA2, SPS19, YAT1 and YAT2) are also located in a downregulated cluster associated with the cytoplasmic organic substance metabolic process (C4). Peroxisomes are involved with long fatty acid degradation and biosynthesis in yeasts (JANG; LIM; KIM, 2014; ZHOU et al., 2016), with lipid metabolism/biogenesis being one of the downregulated gene clusters identified (C10). In contrast, we observe an increased expression in genes associated with nuclear ribogenesis (C3) and metabolic processes (C6 and C12). When compared with the predicted pathway impact, we observe that C3 has 19 of the 23 DEGs associated with ribosome biogenesis pathway (AFG2, DIP2, EMG1, KRE33, POP1, POP7, PWP2, RIO2, RIX7, RNT1, SDO1, UTP13, UTP18, UTP22, UTP4, UTP5, UTP6, UTP8, UTP9), C6 contains 14 of the 31 DEGs associated with purine metabolism (ADE1, ADE12, ADE13, ADE17, ADE2, ADE4, ADE5,7, ADE6, ADE8, ADK1, GUD1, HPT1, IMD2 and IMD4) and C12 contains 13 of the 31 DEGs associated with the biosynthesis of amino acids (ARG1, ASN1, GLY1, HIS1, HIS4, HIS6, HIS7, HOM3, LYS2, LYS4, SER1, SER2 and SHM2). When coupled with the alterations observed in gene sets associated with the mitochondria, our data implies that multiple biological pathways are being regulated to compensate for the stress induced by pCA exposure. Such regulation might be associated with biological processes used by SA1 yeast to survive under adverse conditions.

In addition to gene expression changes, mutations are also a major player in the process of generating resistant strains for biofuel production, be them artificially generated or naturally selected (QI et al., 2014; REID et al., 2011; YANG et al., 2018). These structural alterations can promote changes in the function of genes and proteins in multiple ways (SOSKINE; TAWFIK, 2010), with even small mutations possibly having far-reaching effects (N et al., 2015; ZHANG; CASE; PENG, 2018). By characterizing the profile of single nucleotide variants when compared to the S288C reference strain (which has extensive functional gene annotation data), we are able to predict the impacts of mutations found in the target SA-1 strain genes (ZHANG et al., 2018). Although we recognize that a comprehensive analysis based on comparative genomics of industrial yeasts, which is outside the scope of the present article, would be ideal to characterize the genomic complexity of the SA-1 industrial strain, we expect that extracting the information on the genetic diversity encountered in SA-1 will provide an important layer of information on the mechanisms associated to *pCA* response on this strain.

In order to create a panorama of the potential mechanisms involved in SA-1's tolerance to *pCA* stress, we converted data layers discussed previously (metabolites/phenotypes, transcriptomics and genomics) into a network model and combined those data layers into a unified multi-omics graph model. This type integrative approach is especially relevant to "Big Data" datasets, such as ours, in order to extract comprehensive models that capture subtleties involved in biological regulation that otherwise would be lost if each "omic" was only analyzed independently (CUI; CHENG; ZOU, 2021). This type of analysis has already been successfully used in multiple fields, from biomedical research (HOANG et al., 2019; KHELLA et al., 2021) to biotechnology (LU et al., 2018; WANG et al., 2022). When applied specifically to *S. cerevisiae*, this strategy has also been proven to be crucial in unraveling novel molecular mechanisms associated with gene regulation (MARTÍNEZ-MATÍAS et al., 2021), stress tolerance (KANG et al., 2019) and selection of targets for bioengineering (LIU et al., 2021). By anchoring our network on genes and using their associations (be them positive, negative or neutral) to pathways, phenotypes and genomic variants, we were able to identify two main groups of targets: those associated with ethanol production and those associated with biomass yield. Each of these groups have distinct

profiles, especially in their target pathways, so we will be discussing them separately. In order to facilitate the discussion we will be focusing on the hub genes identified previously, since they are the most representative targets for their respective co-expressed clusters. However, we strongly encourage the exploration of our entire network model made available in the supplementary files (Table S9) by the readers.

When analyzing the first group of genes (ethanol related) of the integrated response model, the TDA10 gene, an ATP-binding protein with unknown function that resembles *E. coli* kinases (DE LA SIERRA-GALLAY et al., 2004; HIGGINS et al., 2003), was downregulated ($\log_2\text{FoldChange}$ -0.72) and had inverse relation with glycine, serine and threonine metabolism and ethanol production. In addition, the TDA10 gene was also the target of a homozygous missense variant in the position 343 of the CDS, which changes the corresponding amino acid from a phenylalanine to a leucine, causing structural changes to the overall protein. We also observed a negative correlation with ethanol production for the ERG10 and IDP2 genes. The former (ERG10) may act in the oxidative stress response (HIGGINS et al., 2003) and its deletion was associated with slower doubling times and susceptibility to high NaCl concentrations (BHATTACHARYA; ESQUIVEL; WHITE, 2018), being a major target for genetic engineering approaches (JIA et al., 2019; KWAK et al., 2017; LIU et al., 2019a). The latter (IDP2) is an isocitrate dehydrogenase that was downregulated in our dataset ($\log_2\text{FC}$ -3.57) and has been linked to small reductions in yeast lifespan (LASCHOBBER et al., 2010) – this gene was also downregulated in mutants susceptible to thermosensitive autolysis and associated with mitochondrial dysfunction (ZHANG et al., 2020).

We also identified 3 genes that had positive correlation with ethanol production: ARO10, GCV1 and CHA1 (Figure 3C). Besides its regulatory role in fermentation (DEED et al., 2019), ARO10 acts in the detoxification of damaged amino acids and resistance to lignocellulosic compounds, such as HMF and furfural (LIU; MA, 2020). We found this gene upregulated ($\log_2\text{FoldChange}$ 1.09) upon exposure to 5 mM of pCA, with a positive correlation to ethanol production. The GCV1 gene, upregulated in our dataset ($\log_2\text{FC}$ 0.82), encodes the T subunit of the mitochondrial

glycine decarboxylase system and increases in expression under multiple types of stress responses in *S. cerevisiae* (ALONSO-MONGE et al., 2001; CHANDLER et al., 2004; DE MELO et al., 2010; MORRISSETTE; ROLFES, 2020); however, the exact role of GCV1 in these scenarios is still not fully understood. Lastly, the CHA1 was slightly upregulated in our dataset (log2FC 0.51) and had positive correlation with the biosynthesis of amino acids and metabolism of glycine, serine and threonine. This gene catalyzes the degradation of L-serine and L-threonine to use them as nitrogen sources and is upregulated in the response to ethanol stress (DONG et al., 2017) and in congo red (GARCÍA et al., 2004).

Both CYB2 and CYT1 are mitochondrial genes that are regulated during changes in the anaerobic metabolic processes and fermentation of glucose (LODI; GUIARD, 1991; ZHANG et al., 2017; ZITOMER; LOWRY, 1992). In our dataset, both genes showed a negative correlation with metabolic pathways (that is, they were downregulated while the pathway was activated). Moreover, they appear to have inverse relations with phenotypic changes: CYB2 has a positive (direct) relation with biomass yield, and CYT1 has a negative (inverse) relation with the ethanol output.

In the second group of genes (biomass related) we identified a total of four genes (SOD1, CTA1, CYB2 and SPO14) with positive influence on the biomass yield and five with negative correlation (IRC7, UTP13, MET3, ADK1 and ADE13). One of the most interesting targets in this group is SOD1, a downregulated gene (log2FC -1.34): it encodes a Cu-Zn superoxide dismutase that has the main role of catalyzing the breakdown of toxic superoxides in the cell (BERMINGHAM-MCDONOGH; GRALLA; VALENTINE, 1988) and is also involved in signaling processes involving oxygen and glucose stimuli (REDDI; CULOTTA, 2013). However, recent studies suggested that the main biological role of these proteins in yeasts is the peroxide signaling and activation of peroxisomes and multiple cell homeostasis pathways (MONTLLOR-ALBALATE et al., 2019). This is in accordance with our findings for both gene expression, pathway impact and co-expressed gene clusters enrichment. The other gene associated with response to reactive oxygen species was CTA1, a downregulated gene (log2FC -2.84) in our dataset. This gene encodes a catalase associated with ROS detoxification in peroxisome and in the mitochondria (PETROVA

et al., 2004), and its activity is relevant in oxidative (KURITA, 2003), acetic acid (GIANNATTASIO et al., 2005) and heat (DAVIDSON et al., 1996) stress responses. These genes showed positive associations with the activity of the peroxisomal pathway, which is one of the most affected by the stress induced by pCA in SA1 yeasts. While SOD1 did not appear as a hub in the selected gene clusters, it did act as a major interactor for cluster 8, which is associated with regulatory and cell homeostasis pathways (Table S5). However, CTA1 is a major interaction hub for C4, being enriched for genes related to the metabolism of organic substances in the cytoplasm. Additionally, the SPO14 (log2FC -0.82) gene was associated with changes in the cell cycle regulation (Honigberg et al. 1992) and regulation of lipid metabolism (CARMAN; HAN, 2011; HENRY; KOHLWEIN; CARMAN, 2012) in *S. cerevisiae*. In addition to its role in cell cycle, SPO14 was also a hub gene for cluster 10, which is enriched in genes for lipid metabolic process.

As for the genes that had negative correlations with the biomass yield, the IRC7 gene seems to be a target with multiple associated conditions. Besides its transcriptional behavior (log2FC 0.60) in SA1 yeasts, when exposed to pCA, this gene is involved in the production of thiol compounds (RONCORONI et al., 2011) and yeast survivability using cysteine as nitrogen source (SANTIAGO; GARDNER, 2015). This gene was the most affected by our analysis of variants, accumulating a large amount of moderate-to-high impact variants in heterozygosity, but none in homozygosity, suggesting that one of the alleles might be severely impaired. This corroborates an analysis of wine fermenting yeasts, which showed that several *S. cerevisiae* strains carried inactivating mutations for one or both alleles of IRC7 (CORDENTE et al., 2019), reducing the overall enzymatic activity of this protein. Other study showed that the over-expression of IRC7 also resulted in the increased production of hydrogen sulfide (SANTIAGO; GARDNER, 2015), a volatile sulfur compound that has been linked to increased longevity in *S. cerevisiae* (HUANG et al., 2017). Lastly, we also identified 3 genes that showed a positive correlation with purine metabolism: MET3 (log2FC 0.73), ADK1 (log2FC 0.50) and ADE13 (log2FC 0.82). Upregulated in our dataset (log2FoldChange 0.73), MET3 is an ATP sulfurylase involved in sulfate and methionine metabolism (MENDOZA-CÓZATL et al., 2005, p.), which was upregulated during hypoxia (KITAGAKI; TAKAGI, 2014). Moreover, the over-expression of ADE13 may

increase fermentation efficiency under acetic acid stress (ZHANG et al., 2019), while ADK1 appears to be activated in response to sulphuric acid (DE LUCENA et al., 2015) and to heat stress (AUESUKAREE et al., 2009). These three genes were also associated with clusters enriched in genes linked to the regulation of metabolic processes.

By cross-referencing the findings from the multi-omics para-coumaric response model generated in our study with the comparative genomics analysis for SA-1 strain, we were able to further filter the targets identified. This allowed us to not only identify which targets are unique to SA-1's response from the ones associated with generic stress response to industrial conditions (EIGENFELD; KERPES; BECKER, 2021; GARAY-ARROYO et al., 2003; KIM et al., 2013; MUELLER et al., 2020), but also identify which events are uniquely associated with this particular strain and which are not relevant and/or unique to SA-1. For example, a recent study showed that sequence and structural alterations in IRC7 gene were described as being a common feature among brazilian industrial strains (JACOBUS et al., 2021). This is reflected in our own study, with IRC7 being flagged as a hub gene for response of para-coumaric stress but without any distinguishing features for the SA-1 strain from our comparative genomics data. On the other hand, SPO14, another hub gene identified on our dataset that has been linked to stress response in *S. cerevisiae* (BARMAN et al., 2018; ZHANG et al., 2016), showed several distinguishing features that differentiate the protein from SA-1 from other brazilian industrial strains.

CONCLUSION

Our results suggest that p-Coumaric acid (pCA) stress may induce higher cellular activity in SA-1 yeasts, with increased glucose uptake, CO₂ and ethanol production being the major indicators obtained from HPLC data. In accordance, we also observed a decrease in biomass yield and overall dry mass, which implicates the existence of some type of disturbance in the cell homeostasis. We also demonstrated that pCA stress can cause an overall activation of metabolic and biosynthesis pathways, which are also followed by increased rRNA biogenesis. Downregulation of several mitochondrial and peroxisomal-associated pathways may also be an indicator of cellular damage caused by the exposure to pCA; our data suggests that SA-1 yeasts

have yet-to-be-explored molecular mechanisms that allow them to circumvent triggers that lead to programmed cell death. At the gene level, we identified multiple genes that could be novel and/or interesting targets for bioengineering. Our results highlight the importance of an integrated approach for target identification and association with phenotypes of interest for industrial applications. By using network-enhanced gene cluster detection, we identified the genes that could be the most influential in their biological vicinity. These “hub genes” are prime targets for genetic engineering approaches, as they are the ones with the highest impact on their sphere of influence and most-likely to produce deep alterations in the associated biological process within the gene community. Although exploratory in nature, the data presented in this study contributes to understand the characteristics of *pCA*-induced stress in *S. cerevisiae* and to deepen the knowledge on mechanisms used by industrial yeast strains that can thrive under high-stress conditions, such as the exposure to lignocellulosic inhibitors.

Taken together, our results show that the biological mechanisms used by SA-1 yeasts to survive under the influence of lignocellulosic inhibitors is much more intricate than previously understood. Multiple biological pathways, which sometimes have opposite effects when analyzed individually, are intertwined in a complex balance that allow these yeasts to thrive even when exposed to high levels of stress. A systemic analysis is essential to understand the nuances involved in such interactions, with several information sources and analyses being integrated into a single model that can reflect multiple levels of biological data. This is especially relevant for researches in economically-driven or similar fields, such as bioethanol production and other industrial capacities, where the ability to select targets for bioengineering approaches that maximize the desired effect (e.g. improving ethanol production) while minimizing undesired side-effects (e.g. affecting unrelated pathways and/or other phenotypes) can be of paramount importance to gain a competitive edge. By using our network model as a frame of reference to develop strains that are more robust to the effects of inhibitory compounds, such as *pCA*, we hope to drive innovation towards a more robust yeast strain that is capable of improved efficiency under the strenuous conditions imposed by industrial fermentation vats.

CHAPTER 2 - INDYdb: An integrative comparative genomics database for industrial yeast strains

ABSTRACT

Integrative databases are a rich source of information in the Big Data era. The ability to integrate data sources and provide centralized access to features in a user-friendly manner is of paramount importance to push the boundaries of data-driven science and research. This concept is particularly important to industrial and biotechnology-associated fields, where the ability to quickly gather, analyze and extract meaningful biological data can be the difference between reaching a paradigm-shifting breakthrough or lagging because of a patent held by a competitor. The current state for industrial yeasts databases is worrying, with the main functional annotation database, the *Saccharomyces* Genome Database, having information on only 25 industrial strains. With over 1000 genomes already deposited in GenBank and the increased need for comparative genomics studies for industrial applications, it is of foremost importance the existence of a centralized database for industrial and commercially relevant yeast strains. Here, we present INDYdb, an integrative database for comparative genomics of industrial yeast strains. Our database provides structural and functional annotation for genomes available in public databases, compiling the information in a SQL database that allows relational comparisons of genes across multiple strains. Furthermore, our database allows researchers to perform complex biological investigations such as: identifying characteristics that are unique and/or shared between individuals or groups of strains; establishing evolutionary relationships among strains and generating hypotheses based on a biological question. To achieve this, we developed a fully automated, self-contained, and scalable pipeline compatible with current HPC platforms, requiring minimal human input and capable of going from the raw genome to full annotation in under 6 hours. As a proof-of-concept, we applied the INDYdb pipeline to 26 different strains of *Saccharomyces* that are currently in industrial use in both bioethanol and food production fields.

MATERIAL AND METHODS

Yeast strain and genome assembly selection

In order to be included in INDYdb strains will be selected based on a number of criteria which include:

- Genome assembly must be publicly available at GenBank
- Currently in use in either industry or research
- Have a distinct ID (i.e., strain name) that allows traceability
- Associated publication (with DOI) or GenBank study accession number
- Disclosed geographical location of the sample collection site

Upon selection, the most recent genome assembly (identified by the GCA prefix) version available at GenBank will be downloaded to our local computational cluster and processed using the structural and functional annotation pipelines (described below). In the case of strains with a reference annotation (such as the case of S288C), the reference assembly (identified by the GCF prefix) will be used instead of the raw assembly.

Structural gene annotation

Raw genome sequences will be processed using the LiftOff (LO, SHUMATE; SALZBERG, 2021), a pipeline developed with the specific purpose of accurately mapping GFF/GTF annotations from a reference assembly (in our case, S288C) to target assemblies of the same or closely-related species (Industrial *S. cerevisiae* strains and *S. paradoxus*). Using LO's GFF output, we will extract predicted gene, transcript and protein sequences using GFFRead (PERTEA; PERTEA, 2020), with the following parameters: -F, -V, -B, -H, -E, -M, -K and -Q. This will allow us to prioritize full length protein coding transcripts, while discarding partial, redundant and truncated transcripts. Structural gene annotation quality will be assessed using the percentage of BUSCOs for saccharomycetes (saccharomycetes_odb10) identified (SIMÃO et al., 2015) in the protein sequences extracted by GFFRead. A grid search parameter tuning algorithm was implemented in the pipeline, with each parameter being tuned with 2 additional values in both up or down directions (bringing to a total of 5 values searched per parameter), and the change ratio is directly related to their power, for example: 0.5 becomes [0.3, 0.4, 0.5, 0.6, 0.7]; 5 becomes [3, 4, 5, 6, 7] and

50 becomes [30, 40, 50, 60, 70]. The optimal parameter combination is calculated based on the highest BUSCO score (in percentage of complete and single copy genes) achieved. A minimum value of 95% of BUSCOs identified with complete and single-copy status must be achieved for the annotation to receive a passing grade.

Functional gene annotation

For genome assemblies without a reference annotation, we will perform functional annotation and gene prediction using DeNSAS (De Novo Sequence Annotation System, LABIS, 2021). This is a reference-free pipeline developed in-house that uses a combination of public databases such as PFAM (MISTRY et al., 2021), RFAM (KALVARI et al., 2021), UniProt (BATEMAN et al., 2021), InterPro (BLUM et al., 2021), Merops (RAWLINGS et al., 2018) and RefSeq (O'LEARY et al., 2016) to characterize predicted protein coding sequences according to their most likely gene family, protein domains and gene ontology categories. In addition to characterizing the predicted gene in accordance with these features, DeNSAS also provides a description with the full gene name and unique ID in case the predicted sequence closely matches a known / existing gene in the database. The advantage of DeNSAS over existing pipelines, such as OrthoFinder (EMMS; KELLY, 2019) or CRB-Blast (AUBRY et al., 2014), is that DeNSAS does not require an existing reference strain to guide the annotation process, allowing the accurate characterization and identification of potentially novel genes or duplication events.

Annotation benchmark

Benchmarking of the combined LiftOff (LO) + DeNSAS annotation pipeline was done comparing the predicted genes against the ones predicted by 3-step (ab initio, re-annotation and refinement) Maker v3.01.03 pipeline (CAMPBELL et al., 2014) and annotated using CRB-Blast. Ab initio gene predictors (Augustus (STANKE et al., 2006), GeneMark (BORODOVSKY; LOMSADZE, 2011) and SNAP (KORF, 2004)) were trained using reference S288C strain annotations. Reference strain protein and transcript sequences will also be included as external evidence to be used by Maker 3. In both cases (LiftOff and Maker3 annotations), the predicted annotation was compared with the reference annotation using GFFCompare tool (PERTEA; PERTEA,

2020), this comparison was done for the S288C, CEN.PK1137-D and *S. paradoxus* reference annotations (RefSeq).

Data integration and Database assembly

Structural and Functional annotation data was be compiled and integrated into a single database using customized python 3 (VAN ROSSUM G; DRAKE FL., 2019) scripts, which are available upon request, which generate an SQLite3 indexed relational database that allows users to query gene, transcript and protein sequences in multiple ways. As additional files, that are accessible via an FTP server, we also provide the following:

- Genome assembly in fasta sequences for each strain.
- Complete gene annotation files in GFF format for each strain.
- Gene, transcript and protein sequence fasta files for each strain.
- Multi-sequence alignment files for each gene (SIEVERS; HIGGINS, 2021).
- Newick tree files for each gene (NGUYEN et al., 2015).
- VCF files containing strain-specific variants (PAGE et al., 2016).
- Variant Effect Prediction for gene-level events (MCLAREN et al., 2016).

Web access and API Interface

Access to INDYdb can be done through a public website (<https://indydb.bioinfo.cbmeg.unicamp.br>), which has links to both direct download of the files from the database as well as tutorials guiding the user on how to perform complex queries to the SQL database using a REST API (TARKOWSKA et al., 2018). This API will take in queries in the format of http requests and return the response to the user in the form of a JSON file. These queries can be performed using any combination of the 31 quarriable fields from the database.

Comparative genomics and character tracing

Comparative analysis of the sequences extracted from the annotated genes was performed using OrthoFinder v2.5.4 (EMMS; KELLY, 2017, 2019), using MAFFT (KATO; STANDLEY, 2013) to perform multiple sequence and IQ-Tree (NGUYEN et al., 2015) tree inference. Then, we applied pastML (ISHIKAWA et al., 2019) to perform

character tracing of the annotated phenotypical characteristics of each strain to the species-tree rooted in *S. paradoxus* obtained from OrthoFinder. In order perform statistical analysis of the values found in one particular strain with the remaining population, we applied 1-sample T-test (STUDENT, 1908) with p-values corrected by False-Discovery Rate (BENJAMINI; HOCHBERG, 1995).

RESULTS AND DISCUSSION

Annotation pipeline

The process of adding a new entry to the INDYdb database starts with the selection process of the strain that will be included. Each candidate strain is manually curated by our group through literature review and known applications of that strain. Although we focus on strains associated with industry applications and overall biotechnology research relevance, we also select strains with high relevance to global research landscape (such as S288C, CEN.PK113-7D and the CBS432 strain from *S. paradoxus*), increasing the versatility of our database. Our main goal is to annotate strains which are not found on current databases, such as SGD and Ensembl (HOWE et al., 2020).

Once a new strain is selected, its raw genome is downloaded directly from NCBI's GenBank, then processed through our structural and functional pipeline (Figure 10). By capitalizing on the speed factor provided by LiftOff annotation tool, we implemented a grid search parameter tuning algorithm during the mapping stage in order to achieve the highest possible BUSCO score (in percentage of genes found in complete and single-copy format), with a minimum value of 95% required for inclusion of the strain in the database. On average, this step takes, per strain, from 6 to 8 hours using a 96-core HPC cluster to search all the 78125 possible parameter combinations and guarantee the best possible outcome (with tie-breaking criteria being, in order of importance from high to low, the lowest percentage of missing, fragmented and duplicated BUSCOs, if a tie still persists, we select the prediction that achieved best overall similarity with the reference S288C annotation). In contrast a single run of the Maker3 3-step annotation pipeline, using the same HPC cluster, takes around the same number of hours, requiring multiple runs to find the optimal combination of

parameters to achieve maximum BUSCO score. This makes the Maker3 pipeline, although well-established, impractical for our purposes.

Once the structural annotation pipeline finishes, it automatically moves to the functional annotation step for predicted sequences. Although LiftOff already provides an initial annotation based on the transposed GFF files, we pass the extracted sequences to DeNSAS (De Novo Sequence Annotation System), a pipeline developed in-house that produces reference-free annotations based on similarity score from multiple databases. If the target predicted by DeNSAS is the same as the one predicted by LiftOff, then the gene receives the “Validated” tag in their functional annotation status. If the gene annotated by DeNSAS differs from the one found by LiftOff, it receives the “Mismatch” label. If the gene was not found by DeNSAS (either because it’s a non-coding gene or it didn’t return satisfactory results), it receives the “Not annotated” label. After both the structural and functional annotation pipelines are finished, all the data is then compiled into a single SQLite relational database that uses one-to-many connectors to track the relationships between all information levels: strains, genes, transcripts and proteins (Figure 11). This unified database is then deployed to a docker environment that is accessible via both Web-browser and http-request APIs using our domain: indydb.bioinfo.cbmeg.unicamp.br/INDYdb.

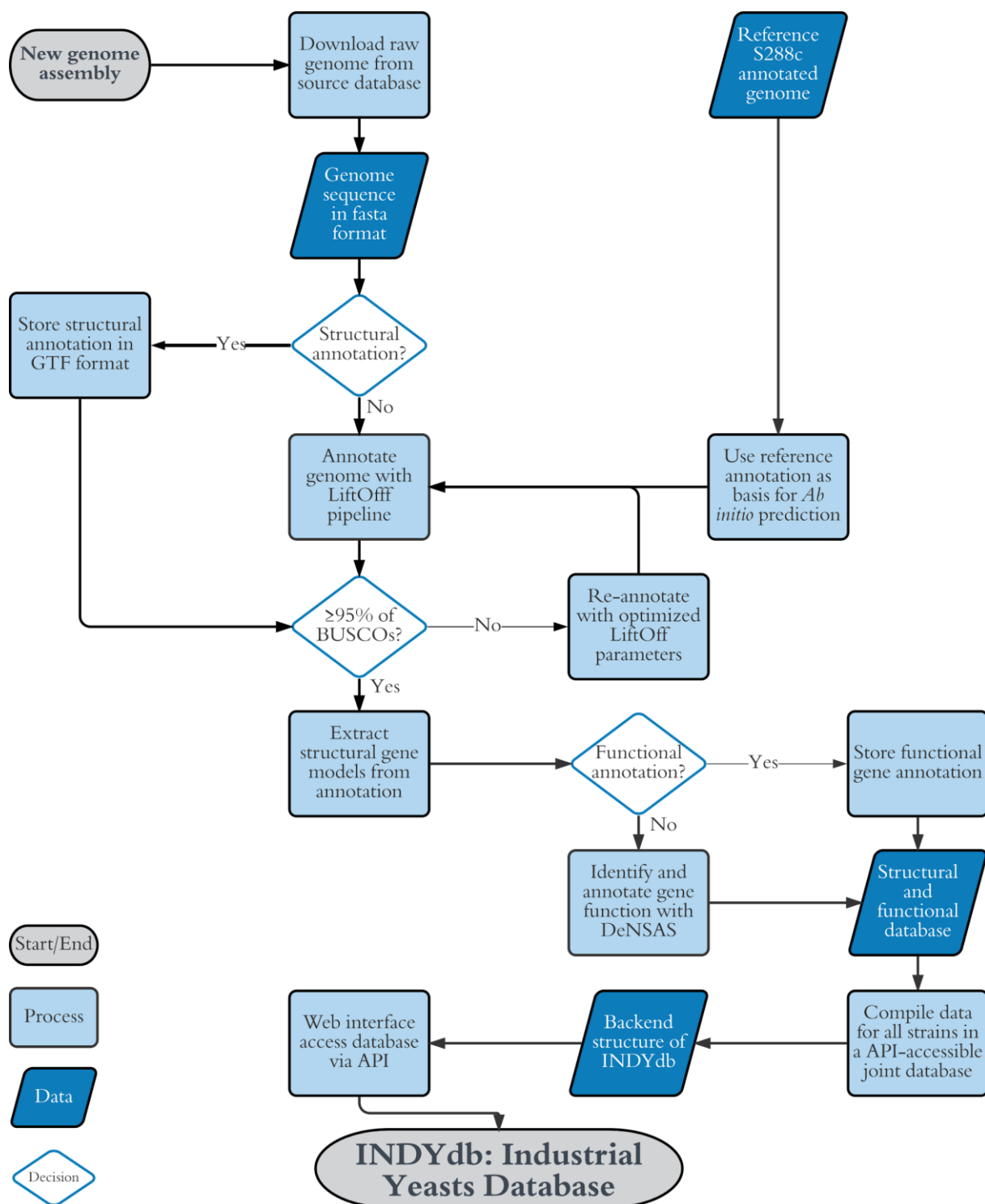


Figure 10. Summary of the computational pipeline used to create INDYdb. The flowchart shows the proposed pipeline used for generating INDYdb, from start to end points and all in-between steps. Start/End steps are shown as gray ovals, processes are represented by light blue rectangles, data is represented by dark blue parallelograms and decision-making steps are represented by white diamond shapes.

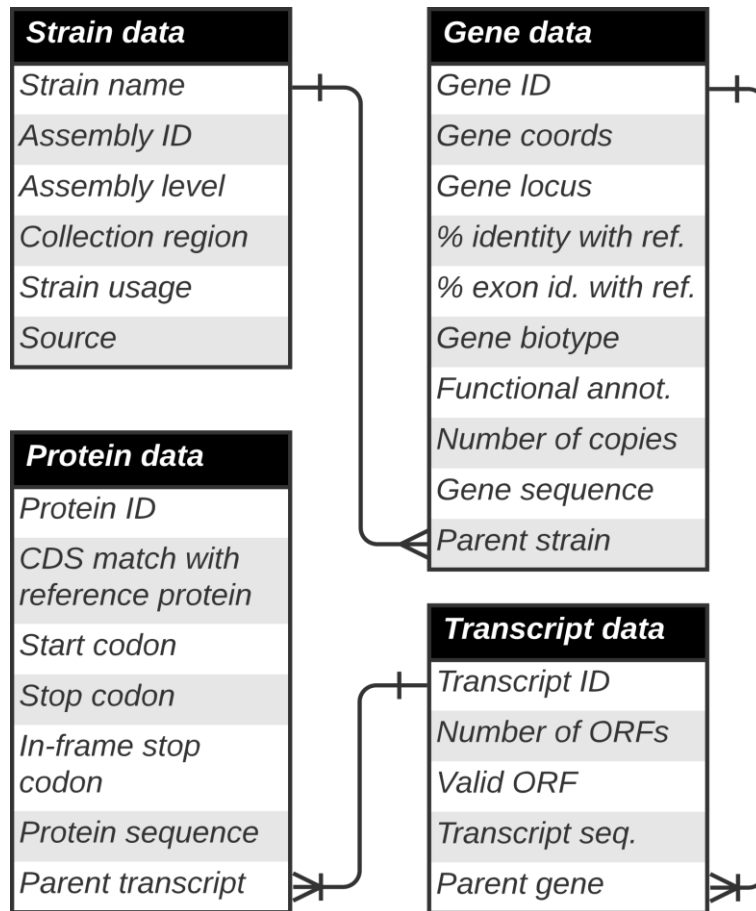


Figure 11. Graphical representation of INDYdb's relational database structure.

INDYdb's internal database structure consists of four different levels of information (represented by the boxes): strain, gene, transcript and protein. These levels are connected with one another via one-to-many connectors (represented by the lines) using a "top-down" approach. Where one strain has multiple genes, each can have multiple transcripts and each transcript can generate multiple protein products. Parent-child between data levels are extracted from the GFF annotation files generated for each strain.

Benchmarking annotation pipeline

In order to evaluate the accuracy of our pipeline when compared to current gold standards, we compared our methodology with the combination of "Maker3 + reciprocal blast" to three strains, two from *S. cerevisiae* (S288C and CEN.PK1337-D) and one from *S. paradoxus* (CBS432) that had reference annotation files that included structural and functional genomic information, with all ab initio models trained in S288C. While the Maker 3 gene prediction pipeline accurately identified ~88.27% of annotated genes from S288C, the LiftOff pipeline annotated ~99.66% of reference sequence genes. Additionally, Maker3 predicted the existence of 109 genes which were not found in the reference S288C annotation, while LiftOff made no such predictions (Figure 12A). The same pattern was observed for the other strains evaluated, with LiftOff outperforming the Maker 3 in both CEN.PK1337-D (96.22% vs. 91.24%) and CBS432 (94.62% vs. 90.92%), and also identifying less genes that had no correspondence with the reference annotation.

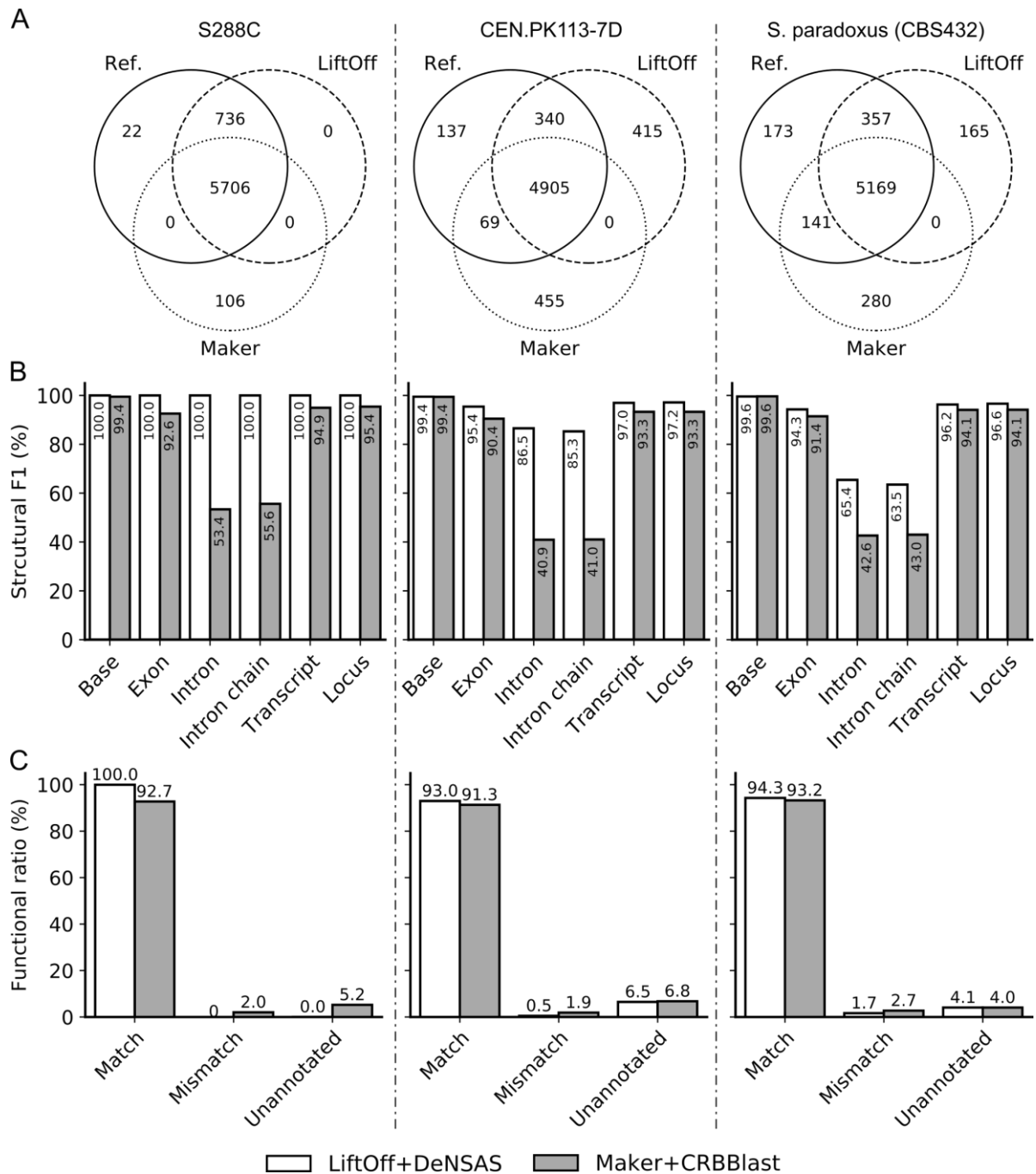


Figure 12. LiftOff combined with DeNSAS shows increased performance for predicting genes in *Saccharomyces* strains. (A) Venn diagram showing the overlap of predicted genes by both LiftOff and Maker3 pipelines when compared to the reference annotation. (B) Barplot showing the structural F1 similarity score in each strain when annotated using LiftOff pipeline (white) and Maker3 (grey). The Y axis show the F1 score in each class (in %) and X axis shows each class evaluated by GFFCompare. (C) Barplot showing the overall accuracy (Y-axis) of prediction by using LiftOff + DeNSAS (white) in comparison with Maker3 + CRB-Blast (gray) pipelines. The X-axis shows the gene classes: Genes where the predicted target matched the reference target were labeled as "Match"; genes which the predicted target differed from the reference were labeled as "Mismatch" and genes from the reference that were missing from the prediction were labeled as "Unannotated".

Then, we compare not only the presence/absence of the genes but also how accurately the position of identified/predicted genes were to their reference annotation counterparts. We used the sensitivity and precision scores generated by GFFCompare to calculate the F1-Score for structural annotation at multiple levels (Figure 12B). Once again, LiftOff outperformed Maker3 for the annotation of genes in each of the tested yeast strains, especially in the case of prediction of intronic chains, which, although uncommon in yeast genomes, are still part of several biologically significant cell activities (CHA et al., 2021; GABUNILAS; CHANFREAU, 2016; RUDAN et al., 2018; SCHREIBER et al., 2015).

Lastly, we compared how accurately our pipeline can functionally annotate the predicted genes predicted by both approaches (Figure 12C). We divided the predicted genes in three distinct groups: Matches (genes where the predicted target has the same annotation as the reference gene), Mismatches (where the predicted targets differ from the reference) and Unannotated (where we could not annotate the predicted target). Once again, LiftOff + DeNSAS outperformed Maker3 + CRB-Blast in all cases, having a higher ratio of matching genes in all strains. Although the two annotation pipelines produced a similar number of unannotated genes, DeNSAS achieved a lower ratio of mismatching genes.

Taken together, these results show that, for the strains analyzed the proposed pipeline for annotation, using LiftOff and DeNSAS, outperforms the Maker + CRB-Blast pipeline, which is considered one of the best practices in current approaches for genome annotation (JUNG et al., 2020; KONG et al., 2019; LANTZ et al., 2018). Furthermore, when comparing with benchmarking studies of gene annotation approaches for multiple clades (SCALZITTI et al., 2020), our pipeline shows increased scores for both structural and functional gene annotation. By using our grid search approach on the algorithms trained on the S288C reference genome, achieved an average structural F1 score of over 99% at nucleotide (base) level and an average of 95% of protein coding genes accurately identified (with an average of 96% protein sequence identity), which is well above the average of 45% structural F1 and 60% protein identity shown in the benchmark study for 5 different algorithms (including Augustus and SNAP, which are part of Maker pipeline). This highlights the importance

of using integrative approaches, such as DeNSAS, to guarantee that annotated genes are the most accurate representation possible (DANCHIN et al., 2018; MUDGE; HARROW, 2016).

Current strains of INDYdb

Currently, our database consists of 25 different strains from *S. cerevisiae* and one strain from *S. paradoxus* (Figure 13A, Table 4). The majority of them (12, ~46% of total entries in the database) are industrial strains currently being used in bioethanol production activities, with 5 of these being used mainly in Asia and 7 being used in South America. Another major group (7 strains, or ~27%) are used in Sake production in Japan (Asia), with lesser representations being strains used in bread production (4 strains, or ~15%) and laboratory strains (2 strains, 7%) and one strain from a wild yeast (1 strain *S. Paradoxus*, CBS432, or 3.5% of the database). When evaluating the accuracy of the prediction using BUSCO, all strains achieved at least a 95% ratio of benchmarking orthologues in complete and single-copy format, with the exception of ZTW1 strain that achieved a score of 94%, however when also considering the BUSCOs in complete and duplicated format all strains achieved at least a 97% score (Figure 13B).

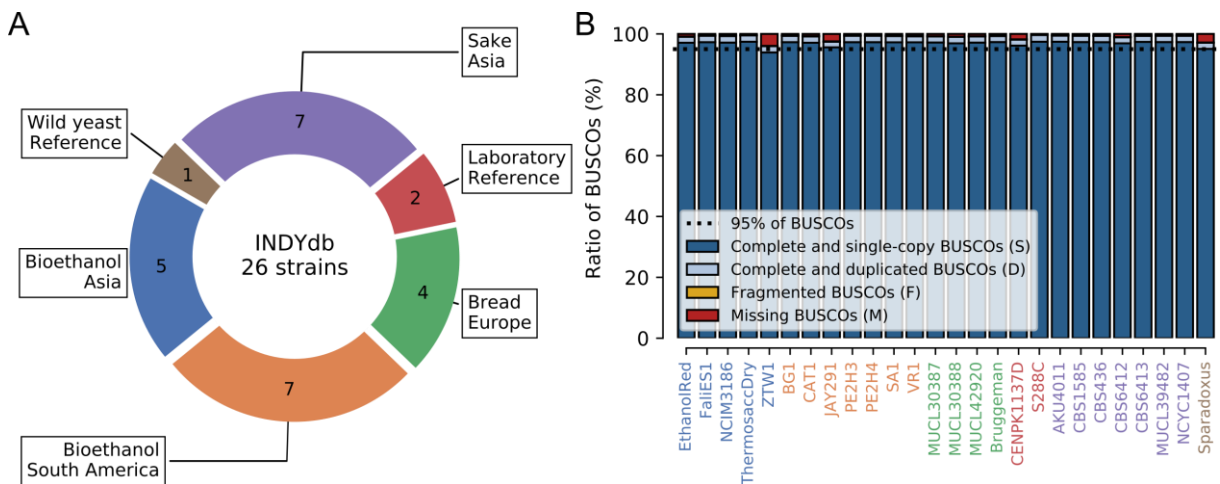


Figure 13. Strains in INDYdb have multiple applications. (A) Circular plot showing the distribution of strains based on their region of origin and current application (in either industry or laboratory). **(B)** Stacked barplot showing the ratio of BUSCOs found in each strain when annotated using LiftOff + DeNSAS pipeline. The Y axis shows the percentage of BUSCOs in each class (colors of the bars) and X axis shows each strain annotated in the database (label colors match the annotation from 1A).

Strain	Assembly	Application	Location	Source
AKU4011	GCA_001738255.1	Sake	Japan	GALLONE et al., 2016
BG-1	GCA_001932575.1	Bioethanol	Brazil	BASSO et al., 2008
Bruggeman	GCA_001738585.1	Bread	Belgium	GALLONE et al., 2016
CAT-1	GCA_001738705.1	Bioethanol	Brazil	GALLONE et al., 2016
CBS1585	GCA_001738375.1	Sake	Japan	GALLONE et al., 2016
CBS436	GCA_001738355.1	Sake	Japan	GALLONE et al., 2016
CBS6412	GCA_001738345.1	Sake	Japan	GALLONE et al., 2016
CBS6413	GCA_001738265.1	Sake	Japan	GALLONE et al., 2016
CEN.PK113-7D	GCA_000269885.1	Laboratory	Netherlands	NIJKAMP et al., 2012
Ethanol Red	GCA_001738615.1	Bioethanol	China	GALLONE et al., 2016
Fal iES1	GCA_001738715.1	Bioethanol	China	GALLONE et al., 2016
JAY291	GCA_000182315.2	Bioethanol	Brazil	ARGUESO et al., 2009
MUCL30387	GCA_001738515.1	Bread	Belgium	GALLONE et al., 2016
MUCL30388	GCA_001738495.1	Bread	Belgium	GALLONE et al., 2016
MUCL39482	GCA_001738235.1	Sake	Japan	GALLONE et al., 2016
MUCL42920	GCA_001738485.1	Bread	Belgium	GALLONE et al., 2016
NCIM3186	GCA_001029075.1	Bioethanol	India	GOUD et al., 2015
NCYC1407	GCA_001738225.1	Sake	Japan	GALLONE et al., 2016
PE-2.H3	GCA_905220325.1	Bioethanol	Brazil	JACOBUS et al., 2021
PE-2.H4	GCA_905220315.1	Bioethanol	Brazil	JACOBUS et al., 2021
S288C	GCF_000146045.2	Laboratory	United States	O'LEARY et al., 2016
SA-1	GCA_004114265.1	Bioethanol	Brazil	NAGAMATSU et al., 2019
Thermosacc Dry	GCA_001738605.1	Bioethanol	China	GALLONE et al., 2016
VR-1	GCA_001738595.1	Bioethanol	Brazil	GALLONE et al., 2016
ZTW1	GCA_000308935.1	Bioethanol	China	ZHANG et al., 2016
<i>S. paradoxus</i>	GCF_002079055.1	Wild yeast	Russia	YUE et al., 2017

Table 4. Summary of strains currently in use by INDYdb. Each entry in the table shows the common name used by the strain, the assembly version used for the annotation pipeline, known main application of the strain, location of the collection site and the publication for the genome sequence.

We annotated, on average, 6262 genes per strain with a ratio of 87% of targets showing positive functional annotation (agreement between structural homology and de novo sequence prediction). The strain with the lowest functional annotation score was *S. paradoxus*'s CBS432, with a verified ratio of 84% of total genes, and the highest scoring was S288C itself, with a ratio of 90% of total genes (Figure 14A). When evaluating the classes for unannotated genes (Figure 14B), we observed that on average only 562 genes per strain remained without a de novo annotation (therefore, the annotation is derived solely from the structural homology), with only ~30% of these targets (an average of ~170 genes per strain) corresponding to protein coding genes, while the remaining 70% of unannotated genes corresponding to non-coding RNAs (~20%), tRNAs (~49%) and rRNAs (~1%), which are notoriously challenging for functional annotation pipelines (EJIGU; JUNG, 2020; PERENTHALER et al., 2019; RAMILOWSKI et al., 2020; SALZBERG, 2019).

In total, our database currently consists of 162036 unique genes which are divided into the following classes: 151872 protein coding; 7089 tRNA; 1977 snoRNA; 385 ncRNA; 184 rRNA; 157 snRNA; 138 misc RNA; 130 antisense RNA; 26 telomerase RNA; 26 RNase MRP RNA; 26 RNase P RNA; 26 SRP RNA. By comparing the sequences of the same gene across multiple strains using multiple sequence alignment, we annotated approximately 3.1 million single nucleotide polymorphism sites (Figure 14C), with an average of ~121 thousand SNP-sites per strain. The strain with the lowest number of variant sites was AKU4011, with 39475 sites detected, and the *S. paradoxus* CBS432 strain had the highest number with 952366 SNP sites detected.

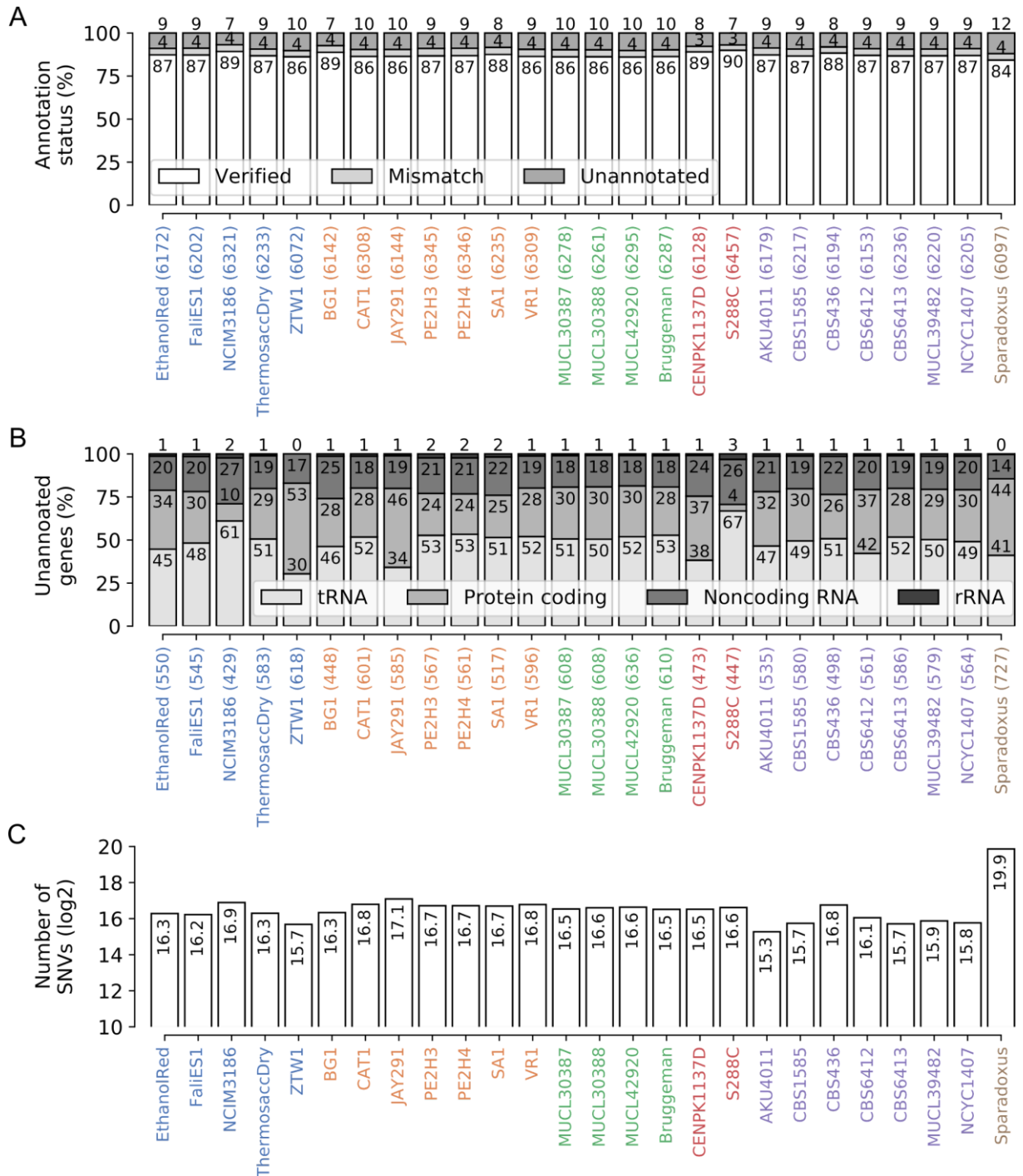


Figure 14. Annotation status profile of current strains in INDYdb. (A) Stacked barplot showing the current strains in INDYdb (X-axis) and the status of the annotation for the genes (Y-axis). There are three major categories: Verified (when structural and functional annotation match), Mismatch (when structural and functional annotation differ) and Unannotated (when there's only structural annotation). **(B)** Stacked barplot showing the distributions of the gene classes for the "Unannotated" genes (Y-axis), with four major classes represented: tRNAs, Protein coding, Noncoding RNAs and rRNAs. **(C)** Barplot showing the number of single nucleotide variant sites (Y-axis) shown for each strain for each strain (X-axis) when compared to the consensus sequence.

Comparative genomics of INDYdb's strains

As an example of potential applications of INDYdb, we did an exploratory comparative genomics analysis to evaluate how the *Saccharomyces* strains currently in use by INDYdb are related to each other. By accessing the “Strain” section of our database (<https://indydb.bioinfo.cbmeg.unicamp.br/Strains.html>), it is possible to directly download the whole genome, GFF annotation files and annotated gene/transcript/protein sequences. Using the extracted protein sequences as input to OrthoFinder pipeline, we were able to easily compare the strains with one another and extract orthology information.

We observed that the tree rooted in the *S. paradoxus* outgroup (Figure 15A), shows a tendency to reflect a combination of both region of origin and application on its branches. When comparing the strain-specific duplication events with at least 50% support on the tree (Figure 15B) we found that *S. paradoxus* had the biggest number of unique events (73 total strain-specific events), followed by S288C (with 50 total events) and NCIM3186 (25 events, the only Indian strain in our database and also the only one that uses sweet sorghum as a fermentation substrate). It is also interesting to note that Brazilian bioethanol strains appear to have a tendency of accumulating almost double the number of duplications (average of 8.16 duplications per strain) when compared to Asian corn-starch fermenter bioethanol-related strains (average of 4.25/strain). When analyzing the ratio of genes placed in orthologous groups (orthogroups) and the ratio of orthogroups containing each strain (Figure 15C), we observe that while nearly all strains had their genes fully included in an orthogroup (average of 99.95%), however there were significant differences in the ratio of orthogroups per strain. Only 90.8% of orthogroups contained genes derived from *S. paradoxus* (which is expected as being the outgroup of this dataset), however we also found that two strains of *S. cerevisiae* showed statistically significantly ($FDR \leq 0.05$) lower scores when compared to the remaining strains: ZTW1 (with genes present in 93.3% of orthogroups) and JAY291 (94.9%). The same pattern can be observed when analyzing pairwise sharing of orthogroups and orthologues (Figure 15D), with these 3 strains exhibiting lower ratios of similarity scores when compared to the other strains of the database.

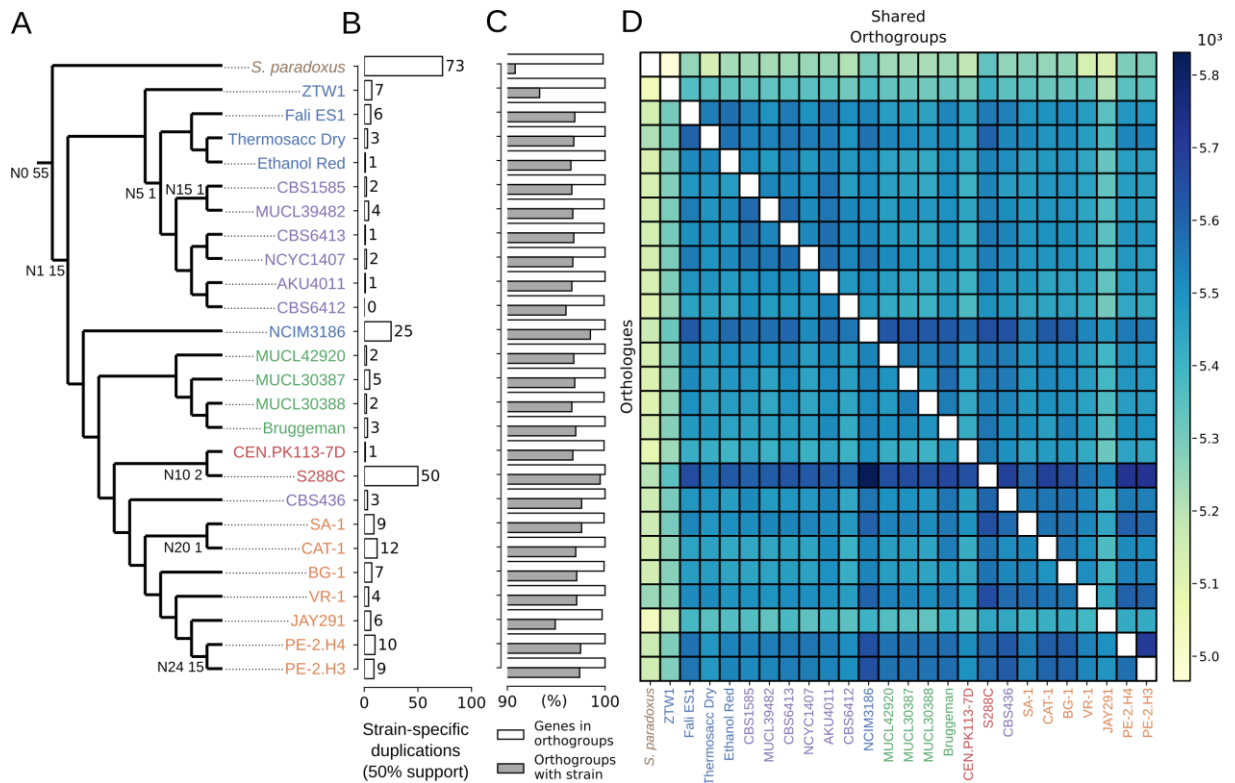


Figure 15. Comparative genomics assessment of yeast strains aggregates strains based on geographical location of collection site. (A) Graphical representation of the newick tree rooted in *S. paradoxus* for each strain (leaves) based on similarity of protein sequences as constructed by OrthoFinder. When numbers are shown on a node (labeled by N followed by a two-digit numeric code), it indicates the number of duplication events that occurred with at least 50% support. **(B)** Barplot showing the number of strain-specific duplication events (X-axis) that occurred with at least 50% support in each strain (Y-axis). **(C)** Barplot showing, for each strain (Y-axis), the ratio (X-axis) of genes placed in an orthogroup (white) and orthogroups containing genes from that particular strain (grey). **(D)** Heatmap showing the pairwise number of shared orthogroups (upper triangle) and orthologous genes (lower triangle) for each strain.

When analyzing the ancestral character tracing over the tree, we observed a strong correlation (joint MAP-MPPA log-likelihood: -35.34) between the tree branches and the geographical collection sites (country) of each strain (Figure 16). However, one particular strain (CBS436) did not cluster together with the other Japanese sake-producing strains and instead seems more related to Brazilian ethanol-producing strains. Taken together, our data suggests that the original center of divergence from *S. cerevisiae* was somewhere in the Far East (near China), then dispersing to other neighboring regions such as Japan and Europe (which then originated the American strains), while the *S. paradoxus* sample (which was collected

in Russia) forms its own branch. These results are in accordance to previously published studies on *S. cerevisiae* domestication origins and dispersion throughout the globe (DUAN et al., 2018; GALLONE et al., 2016, 2018; GODDARD et al., 2010; JACOBUS et al., 2021; LEGRAS et al., 2007).

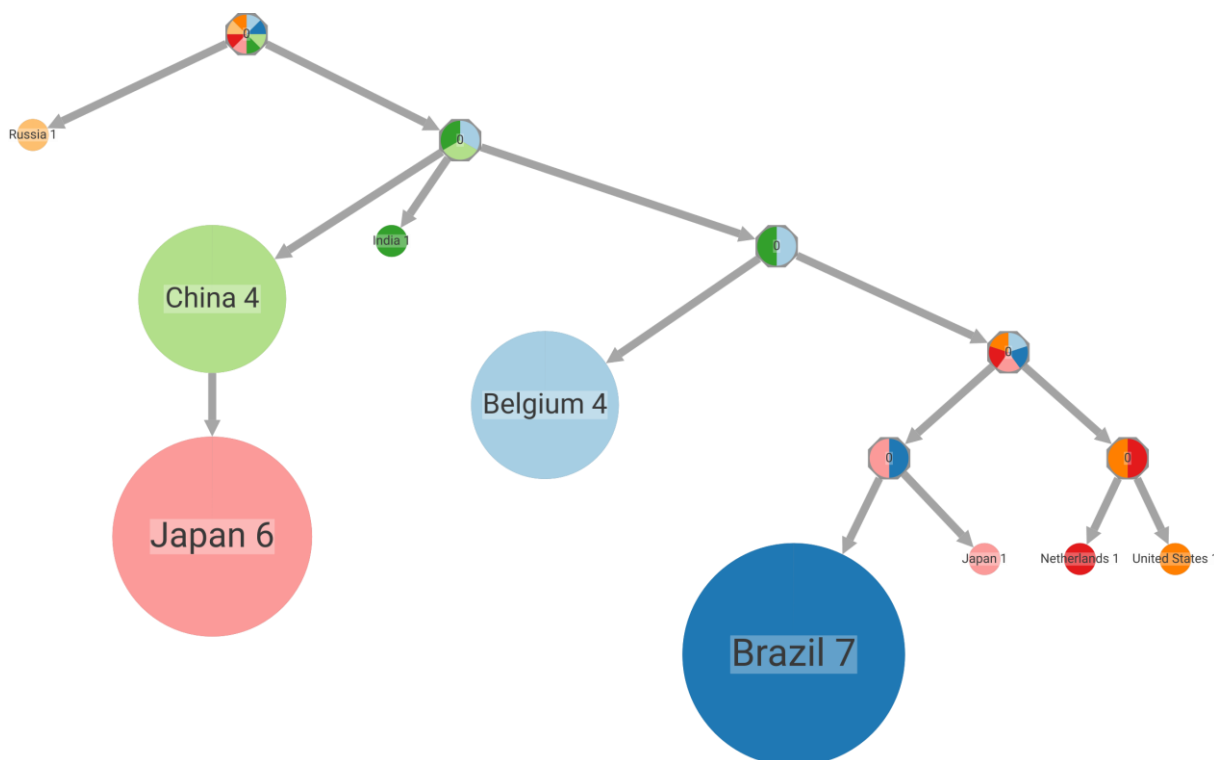


Figure 16. Character tracing reveals geographical ancestry of industrial *S. cerevisiae* strains.

This tree represents a compressed visualization of the character tracing for geographical localization of the collection site based on the species tree generated by OrthoFinder. Each intermediate node represents a division point in the tree branches and the final nodes represent clusters of strains that share the same trait.

Investigating specific gene of interest

Another potential application of INDYdb is the possibility of using our companion API (<https://indydb.bioinfo.cbmeg.unicamp.br/API.html>) as a tool for quick extraction of information from our database using key-value pairs. For example, recently it was shown that the IRC7 gene, a beta-lyase involved in the production of thiols (RONCORONI et al., 2011), has several characteristics unique to brazilian bioethanol strains (JACOBUS et al., 2021). By using a single combination of key-value pairs (gene=IRC7) in our API query, we quickly extracted (query time 594ms) the information for this gene from all 26 current strains in our database. After exporting the

JSON response to a CSV file and converting True/False values to 1/0 numerical data, we applied a simple T-test statistic to perform a three-way comparison of all relevant fields present in our database for brazilian bioethanol strains vs. non-brazilian bioethanol strains vs. non-bioethanol strains. We found statistically significant differences ($FDR \leq 0.05$) for three different classes: Annotation status, percentage of identity with S288C reference and number of single nucleotide variant sites.

When comparing the tree structure for the gene sequences for all strains (Figure 17A, obtainable from <https://indydb.bioinfo.cbmeg.unicamp.br/Genes.html>), we notice that brazilian bioethanol form their own clade, with our de novo sequence annotation system (DeNSAS) also identified that the gene sequences from brazilian bioethanol were predicted as SPAR_F01160, which is the *S. paradoxus* orthologue gene for IRC7 (which showed 90.7% of sequence identity to S288C reference and 96 SNV sites), while genes for other *S. cerevisiae* strains were predicted as similar to S288C's IRC7 (Figure 17B). Furthermore, these strains showed an average of 85.8% identity with the S288C reference and 142 SNV sites (Figure 17C). On the other hand, non-brazilian ethanol strains had an average identity of 99.4% and 2 SNV sites, this is a similar level to non-bioethanol strains with an average identity of 99.7% and also 2 SNV sites per strain. This shows that our pipeline and database, in addition to achieving results similar to the original publication, to those of the original publication, but we were also able to reach the same conclusions as those reached by the authors using a different set of strains. The main difference is that we obtained our results by using a simple online query (e.g., a http GET request) that took milliseconds to return the results followed by a rudimentary statistical analysis that can be done in an excel spreadsheet.

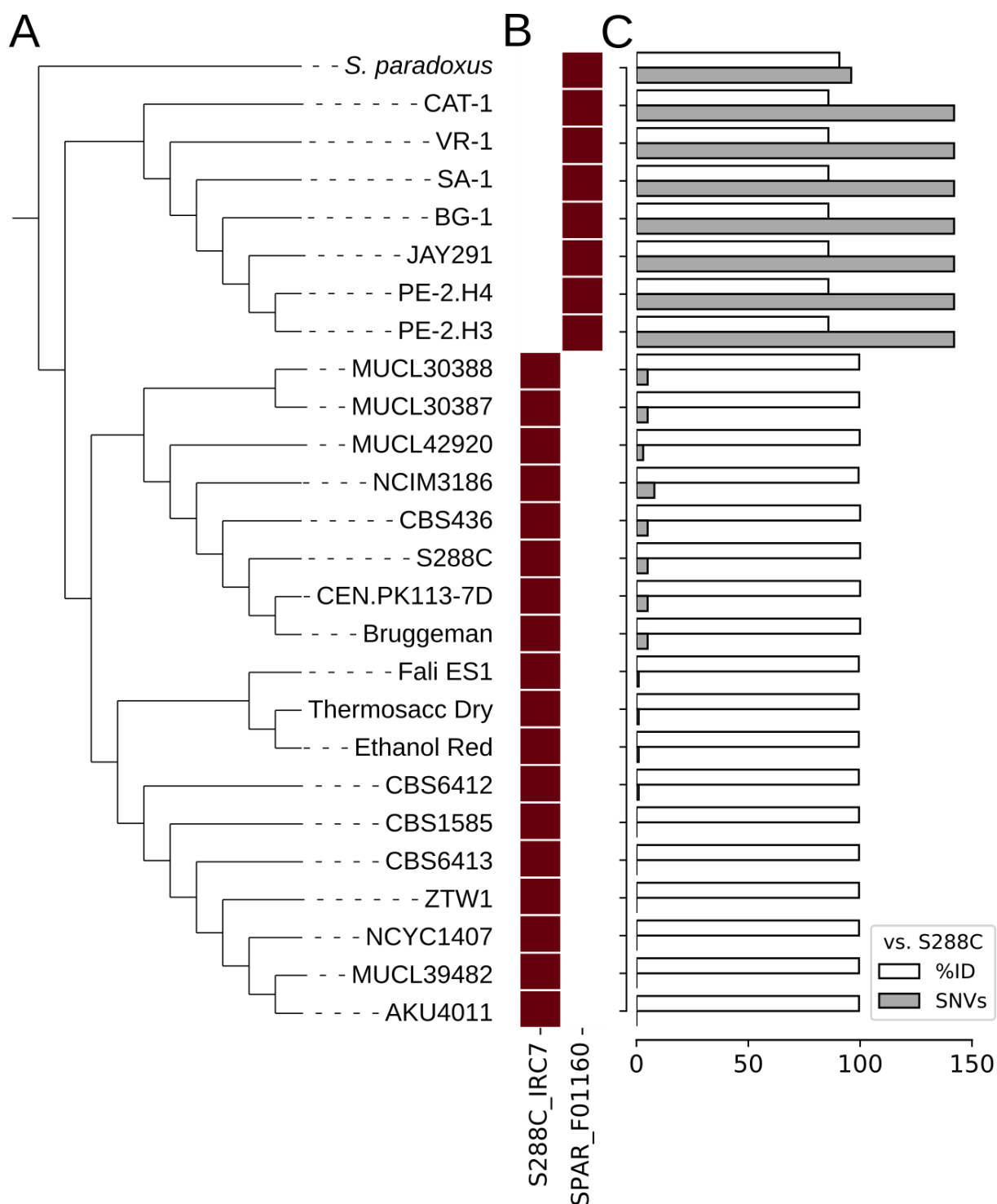


Figure 17. IRC7 genes from Brazilian industrial strains have high similarity to *S. paradoxus* IRC7 orthologue. (A) Tree (rooted in *S. paradoxus* outgroup) showing the graphical representation of the gene sequence similarity for the IRC7 gene. (B) Colormap indicating the results from the De Novo sequence annotation scores, with the gene being similar to either S288C's IRC7 (on the left) or to the SPARA_F01160 (on the right). (C) Barchart showing the overall sequence identity score (white bars) vs. S288C reference strains and the total number of SNV sites found in each of the genes.

Prospection of potential targets based on phenotype

By using a combination of multiple key-value pairs, researchers are also capable of constructing complex queries that reflect biological phenotypes of interest. For example, the question “What are all protein-coding genes related to kinases without a functional ORF in american (either north or south) strains related to ethanol production that also showed less than 95% similarity with the S288C reference?”, could be converted in a query by using the following structure: `type=protein_coding; annotation=kinase; validORF=False; macro_region=america; application=ethanol; %id=<95`. This search, which used 6 different parameters in a combination of text matching, true or false values and numerical thresholds, took an average of 4.81s to query all entries and filter only the targets of interest.

This hypothetical question returned 10 hits in our database, with 5 of those being the CDC4 gene, an F-box protein required for both the G1/S and G2/M phase transitions that forms a complex associated with the ubiquitination of cyclin-dependent kinase (CDK) phosphorylated substrates (FELDMAN et al., 1997; GOH; SURANA, 1999; JACKSON; REED; HAASE, 2006). Upon further evaluation using the same strategy applied to the IRC7 gene, we identified 4 classes that showed statistically significant ($FDR \leq 0.05$) differences between american bioethanol strains and the other strains in the database (Figure 18A): missing stop-codons, absence of a valid ORF, and exon coverage and sequence identity (vs. S288C reference). When analyzing the distributions of these features across the two groups of strains, we observe a distinct pattern between the two groups (Figure 18B). Five of the american bioethanol strains had no stop-codons in their CDC4 sequence, with only 29% of them showing a valid open-reading frame. When comparing the overall exon coverage and identity similarity with the S288C reference, we observe that this subgroup of strains has lower scores in both fields (average of 84% and 82%, respectively). By using multiple alignment of protein sequences anchored on the S288C reference strain (Figure 18C), we can observe that 3 of the strains (SA-1, JAY291 and BG-1) show a truncation of their terminal region, 2 other strains (CAT-1 and VR-1) have significant differences in the sequence of their terminal region and only 2 strains (the two haplotypes from PE-2) showed similarity to S288C.

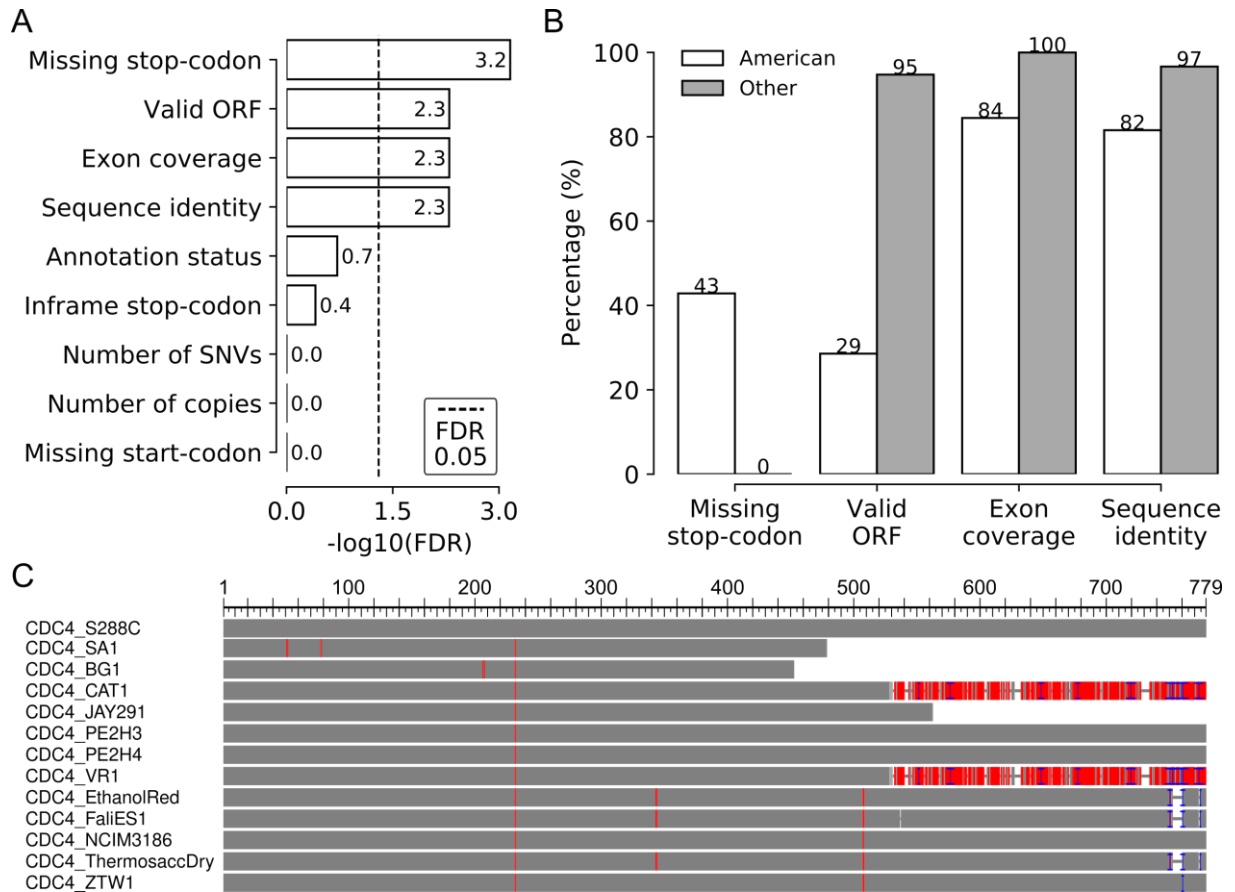


Figure 18. American bioethanol-related strains exhibit a higher proportion of altered genomic features for the CDC4 gene. (A) Barplot showing the FDR values (X-axis in $-\log_{10}$ scale) for the comparison values found in each genomic feature category (Y-axis) of the american population vs. other bioethanol-related yeast strains. **(B)** Barplot showing the average values (Y-axis) found in each of the statistically significant categories (X-axis) for both the american yeast strains (white) and the remaining bioethanol-related strains (gray). **(C)** Multiple sequence alignment of bioethanol-related strains anchored on the S288C reference strains. Gray bars represent the alignment of the protein sequence and the red bars show regions where the target strain sequence differs from the reference.

When we take into account that brazilian industrial yeast strains that are inoculated into fermentation vats are known to suffer displacement from wild yeasts during the ethanol production process (ARGUESO et al., 2009; DA SILVA-FILHO et al., 2005; DELLA-BIANCA et al., 2014; JACOBUS et al., 2021; LUCIO LOPES FERMENTEC; CRISTINA DE LIMA PAULILLO, 2015) and the description of the function of CDC4, one could infer an association between the duplication efficiency of this particular group of yeasts (brazilian industrial strains), the different structural features found in CDC4 and the overall displacement of said strains by yeasts that have a functional CDC4 (a.k.a. “Wild-type”). While this is merely speculation on the

part of the authors of this work, it serves as an example of possible biological questions that can be raised using INDYdb.

CONCLUSION

By using a combination of high-performance computing clusters and a pipeline that uses automated decision-making based on well-defined metrics, we were able to develop an approach that is capable of quickly and accurately annotating publicly available genomes from *S. cerevisiae*. These annotations are then compiled into a relational database that allows direct comparison of the genes among all annotated strains.

These comparisons open the door for a plethora of applications for researchers in multiple fields. From evaluating the relationships of the strains on a global level to identifying potential novel targets for bioengineering, our database provides the scientific community with a rich public repository of information that can be explored via multiple approaches. When we combine our database with an efficient information delivery service in the form of an API that queries our internal SQL database, we further push the boundaries of what kind of questions can be asked by researchers and take advantage of modern big data strategies that are already in use in other fields. For example, future implementations of INDYdb will feature text-to-sql generators that can convert natural language to sql-like queries, making it easier for researchers to access and parse large datasets while focusing on biological questions at hand.

In summary, we show that relational databases can be a major asset for researchers interested in yeast biotechnology. By using a combination of efficient methods for extracting information from large datasets and integrative analysis of said data, it is possible to perform complex questions to our database that can provide valuable insights on the researcher's interest (be it a target gene, a strain or a phenotype), which, in turn, expedites the innovation process for both academic and industrial purposes.

FINAL REMARKS

This thesis was divided into two main segments. In the first chapter, we explored the effects of p-Coumaric acid (pCA) on an industrial brazilian strain of *Saccharomyces cerevisiae* (SA-1) that is highly resistant to lignocellulosic inhibitors. By using a combination of gene expression data, metabolite components measurement, nucleotide variant sites and comparative genomics, we were able to create a comprehensive panorama of the changes incurred by SA-1 yeasts and how this particular strain is capable of responding to stress induced by pCA, which were then summarized in a graph-based network that can further explored by researchers with various interests. Furthermore, we were also able to use network centrality analysis to extract which genes act as interaction hubs inside co-expressed gene clusters and could be potential targets for bioengineering projects.

In the second chapter, we presented a database, called INDYdb, that was originally created out of a necessity to answer complex comparative genomics questions of industrial yeast strains. By taking advantage of genomes deposited in public databases, modern high-performance computing clusters and state-of-art genome annotation algorithms, we developed an integrated and scalable pipeline that is capable of quickly and accurately annotate new genomes and integrate that information into an existing self-contained SQL database structure that is designed with scalability and expansion capacity in mind. In order to access the data contained in our database, we developed two companion tools that allow human interface to our repository. The first is a fully interactive website that allows the user to manually curate the data for individual genes and strains, download portions or the entirety of the database directly and perform simple stand-alone queries. The second tool, which is considerably more powerful, is an integrated REST API that transforms GET requests into SQL queries, parses then into the internal database and returns them as JSON structured-files, which can then be loaded into data analysis platforms such as python notebooks or even excel spreadsheets. By using combinations of any of the 31 queryable fields, these questions can range from simple data retrievals of a single gene, to complex biological scenarios that reflect phenotypes of interest. This database will certainly prove to be an important asset to scientists currently engaged in biotechnology research in either public or private sectors.

Taken together, the results presented here as by the two distinct chapters, although exploratory in nature, highlight the importance and relevance of integrative biology approaches applied to biotechnological research. Both the multi-omics network model for SA-1's response to p-Coumaric acid exposure and the INDYdb relational comparative genomics database are important resources that were made available for the research community. This is especially relevant in the case of this thesis, since it is inserted in the context of Work Package 3 from FAPESP-BBRSC collaboration project (2015/50612-8) entitled "An integrated approach to explore a novel paradigm for biofuel production from lignocellulosic feedstocks". The data presented here will synergize and fuel on-going projects from the remaining Work Packages, specially those produced by the WP2, as engineered *S. cerevisiae* strains will be tested for growth and ethanol productivity on substrates generated from the pre-treatment of cellulose with the processed endoglucanase cocktails characterized in WP2, with the aim of developing a fully integrated process.

Lastly, it is also important to highlight the projections of the work presented here for the future of biotechnology research, this is especially true for the database contained here. Our initial deployment of INDYdb served as a proof-of-concept for the capabilities of these types of databases in pushing innovation, however this project can be improved upon in several ways. The first, and most obvious, is increasing the number of strains currently available and developing methods to make the inclusion of new strains more dynamic in nature. The second is pushing the development of computational tools that make queries to such databases more efficient, either by improving their speed by using better algorithms or ease-of-use by including natural language to sql-query converters. Third is the inclusion of novel analysis tools available to the end-user via friendly graphical interfaces on the website. Last, but not least, is the possibility of converting several of the tools in INDYdb to be presented "as-a-service" to external sources, which could bolster even further the impact of our database. Each of these facets can be forked into a different project, ensuring the continuity of INDYdb as a stable data source. Although the implementation of INDYdb presented here was the first in its lineage, it certainly will most certainly not be the last.

REFERENCES

- ADEBOYE, P. T. et al. Catabolism of coniferyl aldehyde, ferulic acid and p-coumaric acid by *Saccharomyces cerevisiae* yields less toxic products. **Microbial Cell Factories**, 2015.
- ADEBOYE, P. T.; BETTIGA, M.; OLSSON, L. The chemical nature of phenolic compounds determines their toxicity and induces distinct physiological responses in *Saccharomyces cerevisiae* in lignocellulose hydrolysates. **AMB Express**, v. 4, n. 1, 2014.
- ADEBOYE, P. T.; BETTIGA, M.; OLSSON, L. ALD5, PAD1, ATF1 and ATF2 facilitate the catabolism of coniferyl aldehyde, ferulic acid and p-coumaric acid in *Saccharomyces cerevisiae*. **Scientific Reports 2017 7:1**, v. 7, n. 1, p. 1–13, 16 fev. 2017.
- AL NAQBIA et al. The Impact of Innovation on Firm Performance: A Systematic Review. **International Journal of Innovation**, v. 14, n. 5, 2020.
- ALMEIDA, J. R. M. M. et al. Increased tolerance and conversion of inhibitors in lignocellulosic hydrolysates by *Saccharomyces cerevisiae*. **Journal of Chemical Technology and Biotechnology**, v. 82, n. 4, p. 340–349, 2007.
- ALONSO-MONGE, R. et al. Hyperosmotic stress response and regulation of cell wall integrity in *Saccharomyces cerevisiae* share common functional aspects. **Molecular Microbiology**, v. 41, n. 3, 2001.
- ALPER, H. et al. Engineering yeast transcription machinery for improved ethanol tolerance and production. **Science**, v. 314, n. 5805, p. 1565–1568, 2006.
- ARGUESO, J. L. et al. Genome structure of a *Saccharomyces cerevisiae* strain widely used in bioethanol production. **Genome Research**, v. 19, n. 12, p. 2258–2270, 1 dez. 2009.
- ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nature Genetics**, v. 25, n. 1, p. 25–29, maio 2000.
- AUBRY, S. et al. Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two Independent Origins of C4 Photosynthesis. **PLoS Genetics**, 2014.
- AUESUKAREE, C. et al. Genome-wide identification of genes involved in tolerance to various environmental stresses in *Saccharomyces cerevisiae*. **Journal of Applied Genetics**, v. 50, n. 3, 2009.
- AYER, A.; GOURLAY, C. W.; DAWES, I. W. Cellular redox homeostasis, reactive oxygen species and replicative ageing in *Saccharomyces cerevisiae*. **FEMS Yeast Research**, v. 14, n. 1, p. 60–72, 1 fev. 2014.
- BACCOLO, G. et al. Chapter One - Mitochondrial Metabolism and Aging in Yeast. Em: LÓPEZ-OTÍN, C.; GALLUZZI, L. (Eds.). **International Review of Cell and Molecular Biology**. Mitochondria and Longevity. [s.l.] Academic Press, 2018. v. 340p. 1–33.

BARANOWSKI, J. D. et al. Inhibition of *Saccharomyces cerevisiae* by naturally occurring hydroxycinnamates. **Journal of Food Science**, 1980.

BARMAN, A. et al. Phospholipases play multiple cellular roles including growth, stress tolerance, sexual development, and virulence in fungi. **Microbiological Research**, v. 209, p. 55–69, 1 abr. 2018.

BASSO, L. C. et al. Yeast selection for fuel ethanol production in Brazil. **FEMS yeast research**, v. 8, n. 7, p. 1155–1163, nov. 2008.

BATEMAN, A. et al. UniProt: The universal protein knowledgebase in 2021. **Nucleic Acids Research**, v. 49, n. D1, 2021.

BAXEVANIS, A. D.; BATEMAN, A. The importance of biological databases in biological discovery. **Current Protocols in Bioinformatics**, v. 2015, 2015.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 57, n. 1, 1995.

BERMINGHAM-MCDONOGH, O.; GRALLA, E. B.; VALENTINE, J. S. The copper, zinc-superoxide dismutase gene of *Saccharomyces cerevisiae*: cloning, sequencing, and biological activity. **Proceedings of the National Academy of Sciences of the United States of America**, v. 85, n. 13, 1988.

BHATTACHARYA, S.; ESQUIVEL, B. D.; WHITE, T. C. Overexpression or deletion of ergosterol biosynthesis genes alters doubling time, response to stress agents, and drug susceptibility in *Saccharomyces cerevisiae*. **mBio**, v. 9, n. 4, 2018.

BLAZI, L. E. et al. Adaptation Strategy to Increase the Tolerance of *Scheffersomyces stipitis* NRRL Y-7124 to Inhibitors of Sugarcane Bagasse Hemicellulosic Hydrolysate Through Comparative Studies of Proteomics and Fermentation. **BioEnergy Research**, v. 15, n. 1, p. 479–492, mar. 2022.

BLUM, M. et al. The InterPro protein families and domains database: 20 years on. **Nucleic Acids Research**, v. 49, n. D1, 2021.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: A flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 2014.

BONO, H. All of gene expression (AOE): An integrated index for public gene expression databases. **PLoS ONE**, v. 15, n. 1, 2020.

BORELLI, G. et al. Positive Selection Evidence in Xylose-Related Genes Suggests Methylglyoxal Reductase as a Target for the Improvement of Yeasts' Fermentation in Industry. **Genome Biology and Evolution**, v. 11, n. 7, 2019.

BORJA, G. M. et al. Metabolic engineering and transcriptomic analysis of *Saccharomyces cerevisiae* producing p-coumaric acid from xylose. **Microbial Cell Factories**, v. 18, n. 1, p. 191, 5 nov. 2019.

BORNMANN, L.; HAUNSCHILD, R.; MUTZ, R. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from

established and new literature databases. **Humanities and Social Sciences Communications**, v. 8, n. 1, 2021.

BORODOVSKY, M.; LOMSADZE, A. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. **Current Protocols in Bioinformatics**, 2011.

BRAUER, M. J. et al. Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. **Molecular Biology of the Cell**, v. 19, n. 1, 2008.

CAMPBELL, M. S. et al. Genome Annotation and Curation Using MAKER and MAKER-P. **Current Protocols in Bioinformatics**, v. 2014, 2014.

CARBONARA, N.; PELLEGRINO, R. The role of public private partnerships in fostering innovation. **Construction Management and Economics**, v. 38, n. 2, 2020.

CARMAN, G. M.; HAN, G. S. Regulation of phospholipid synthesis in the yeast *Saccharomyces Cerevisiae*. **Annual Review of Biochemistry**, v. 80, 2011.

CHA, S. et al. Differential activation mechanisms of two isoforms of Gcr1 transcription factor generated from spliced and un-spliced transcripts in *Saccharomyces cerevisiae*. **Nucleic Acids Research**, v. 49, n. 2, 2021.

CHANDLER, M. et al. A genomic approach to defining the ethanol stress response in the yeast *Saccharomyces cerevisiae*. **Annals of Microbiology**, v. 54, n. 4, 2004.

CHEN, C.; HUANG, H.; WU, C. H. Protein bioinformatics databases and resources. Em: **Methods in Molecular Biology**. [s.l: s.n.]. v. 1558.

CHEN, Y. A. et al. MANTA, an integrative database and analysis platform that relates microbiome and phenotypic data. **PLoS ONE**, v. 15, n. 12 December, 2020.

CHERRY, J. M. et al. *Saccharomyces* Genome Database: The genomics resource of budding yeast. **Nucleic Acids Research**, v. 40, n. D1, 2012.

CHOROSTECKI, U. et al. MetaPhOrs 2.0: Integrative, phylogeny-based inference of orthology and paralogy across the tree of life. **Nucleic Acids Research**, v. 48, n. W1, 2021.

CIAMPONI, F. E. et al. **IndyDB - a comparative genomics database for industrial yeast strains** Repositório de Dados de Pesquisa da Unicamp, , 2022. Disponível em: <<https://doi.org/10.25824/redu/EMXWAY>>

COLA, P. et al. Differential effects of major inhibitory compounds from sugarcane-based lignocellulosic hydrolysates on the physiology of yeast strains and lactic acid bacteria. **Biotechnology Letters**, 2020.

CORDENTE, A. G. et al. Inactivating mutations in Irc7p are common in wine yeasts, attenuating carbonsulfur β -lyase activity and volatile sulfur compound production. **Applied and Environmental Microbiology**, v. 85, n. 6, 2019.

CUI, F.; CHENG, L.; ZOU, Q. Briefings in functional genomics special section editorial: analysis of integrated multiple omics data. **Briefings in Functional Genomics**, v. 20, n. 4, p. 196–197, 1 jul. 2021.

DA SILVA-FILHO, E. A. et al. Yeast population dynamics of industrial fuel-ethanol fermentation process assessed by PCR-fingerprinting. **Antonie van Leeuwenhoek** **2005** **88:2**, v. 88, n. 2, p. 13–23, ago. 2005.

DANCHIN, A. et al. No wisdom in the crowd: genome annotation in the era of big data - current status and future prospects. **Microb Biotechnol**, v. 11, n. 4, p. 588–605, 1 jul. 2018.

DAVIDSON, J. F. et al. Oxidative stress is involved in heat-induced cell death in *Saccharomyces cerevisiae*. **Proceedings of the National Academy of Sciences of the United States of America**, v. 93, n. 10, 1996.

DE LA SIERRA-GALLAY, I. L. et al. Crystal Structure of the YGR205w Protein from *Saccharomyces cerevisiae*: Close Structural Resemblance to *E. coli* Pantothenate Kinase. **Proteins: Structure, Function and Genetics**, v. 54, n. 4, 2004.

DE LUCENA, R. M. et al. Transcriptomic response of *Saccharomyces cerevisiae* for its adaptation to sulphuric acid-induced stress. **Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology**, v. 108, n. 5, 2015.

DE MELLO, F. DA S. B. et al. Static microplate fermentation and automated growth analysis approaches identified a highly-aldehyde resistant *Saccharomyces cerevisiae* strain. **Biomass and Bioenergy**, 2019.

DE MELO, H. F. et al. Physiological and molecular analysis of the stress response of *Saccharomyces cerevisiae* imposed by strong inorganic acid with implication to industrial fermentations. **Journal of Applied Microbiology**, v. 109, n. 1, 2010.

DEED, R. C. et al. The role of yeast ARO8, ARO9 and ARO10 genes in the biosynthesis of 3-(methylthio)-1-propanol from L-methionine during fermentation in synthetic grape medium. **FEMS Yeast Research**, v. 19, n. 2, 2019.

DELLA-BIANCA, B. E. et al. Physiology of the fuel ethanol strain *Saccharomyces cerevisiae* PE-2 at low pH indicates a context-dependent performance relevant for industrial applications. **FEMS Yeast Research**, v. 14, n. 8, 2014.

DOBIN, A. et al. STAR: Ultrafast universal RNA-seq aligner. **Bioinformatics**, v. 29, n. 1, p. 15–21, 2013.

DONG, Y. et al. RNA-Seq-based transcriptomic and metabolomic analysis reveal stress responses and programmed cell death induced by acetic acid in *Saccharomyces cerevisiae*. **Scientific Reports**, v. 7, 2017.

DUAN, S. F. et al. The origin and adaptive evolution of domesticated populations of yeast from Far East Asia. **Nature Communications** **2018** **9:1**, v. 9, n. 1, p. 1–13, 12 jul. 2018.

EIGENFELD, M.; KERPES, R.; BECKER, T. Understanding the Impact of Industrial Stress Conditions on Replicative Aging in *Saccharomyces cerevisiae*. **Frontiers in Fungal Biology**, v. 0, p. 17, 2 jun. 2021.

EJIGU, G. F.; JUNG, J. Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing. **Biology** **2020**, Vol. 9, Page

295, v. 9, n. 9, p. 295, 18 set. 2020.

EMMS, D. M.; KELLY, S. STRIDE: Species Tree Root Inference from Gene Duplication Events. **Molecular Biology and Evolution**, v. 34, n. 12, p. 3267–3278, 1 dez. 2017.

EMMS, D. M.; KELLY, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. **Genome Biology**, 2019.

ENCODE. ENCODE Guidelines and Best Practices for RNA-Seq: Revised December 2016. n. December, p. 1–5, 2016.

ENRIGHT, A. J.; VAN DONGEN, S.; OUZOUNIS, C. A. An efficient algorithm for large-scale detection of protein families. **Nucleic Acids Research**, 2002.

FAVARO, L.; JANSEN, T.; VAN ZYL, W. H. Exploring industrial and natural *Saccharomyces cerevisiae* strains for the bio-based economy from biomass: the case of bioethanol. **Critical Reviews in Biotechnology**, 2019.

FELDMAN, R. M. R. et al. A complex of Cdc4p, Skp1p, and Cdc53p/cullin catalyzes ubiquitination of the phosphorylated CDK inhibitor Sic1p. **Cell**, v. 91, n. 2, 1997.

FENG, X.; ZHAO, H. Investigating host dependence of xylose utilization in recombinant *Saccharomyces cerevisiae* strains using RNA-seq analysis. **Biotechnology for Biofuels**, v. 6, n. 1, 2013.

FLETCHER, E.; BAETZ, K. Multi-Faceted Systems Biology Approaches Present a Cellular Landscape of Phenolic Compound Inhibition in *Saccharomyces cerevisiae*. **Frontiers in Bioengineering and Biotechnology**, v. 8, 2020.

FROLOVA, A.; WILCZYŃSKI, B. Distributed Bayesian networks reconstruction on the whole genome scale. **PeerJ**, 2018.

FUCHS, B. B.; MYLONAKIS, E. Our paths might cross: The role of the fungal cell wall integrity pathway in stress response and cross talk with other stress response pathways. **Eukaryotic Cell**, 2009.

GABUNILAS, J.; CHANFREAU, G. Splicing-Mediated Autoregulation Modulates Rpl22p Expression in *Saccharomyces cerevisiae*. **PLoS Genetics**, v. 12, n. 4, 2016.

GALLONE, B. et al. Domestication and Divergence of *Saccharomyces cerevisiae* Beer Yeasts. **Cell**, v. 166, n. 6, 2016.

GALLONE, B. et al. Origins, evolution, domestication and diversity of *Saccharomyces* beer yeasts. **Current Opinion in Biotechnology**, v. 49, 2018.

GARAY-ARROYO, A. et al. Response to different environmental stress conditions of industrial and laboratory *Saccharomyces cerevisiae* strains. **Applied Microbiology and Biotechnology** **2004 63:6**, v. 63, n. 6, p. 734–741, 9 ago. 2003.

GARCÍA, R. et al. The Global Transcriptional Response to Transient Cell Wall Damage in *Saccharomyces cerevisiae* and Its Regulation by the Cell Integrity Signaling Pathway. **Journal of Biological Chemistry**, v. 279, n. 15, 2004.

GATK. **RNAseq short variant discovery (SNPs + Indels)**. Disponível em: <<https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels->>>.

GENE ONTOLOGY CONSORTIUM. The Gene Ontology resource: enriching a GOLD mine. **Nucleic Acids Research**, v. 49, n. D1, p. D325–D334, 8 jan. 2021.

GIANNATTASIO, S. et al. **Acid stress adaptation protects *Saccharomyces cerevisiae* from acetic acid-induced programmed cell death**. *Gene. Anais...*2005.

GODDARD, M. R. et al. A distinct population of *Saccharomyces cerevisiae* in New Zealand: evidence for local dispersal by insects and human-aided global dispersal in oak barrels. **Environ. Microbiol.**, v. 12, n. 1, p. 63–73, jan. 2010.

GOH, P.-Y.; SURANA, U. Cdc4, a Protein Required for the Onset of S Phase, Serves an Essential Function during G 2 /M Transition in *Saccharomyces cerevisiae*. **Molecular and Cellular Biology**, v. 19, n. 8, 1999.

GOUD, B. S.; ULAGANATHAN, K. Draft Genome Sequence of *Saccharomyces cerevisiae* Strain NCIM3186 Used in the Production of Bioethanol from Sweet Sorghum. **Genome Announcements**, v. 3, n. 4, 2015.

GU, H.; ZHANG, J.; BAO, J. Inhibitor analysis and adaptive evolution of *Saccharomyces cerevisiae* for simultaneous saccharification and ethanol fermentation from industrial waste corncob residues. **Bioresource Technology**, 2014.

GUPTA, M. K.; CHANDRA, P. A comprehensive survey of data mining. **International Journal of Information Technology (Singapore)**, v. 12, n. 4, 2020.

HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J. **Exploring Network Structure, Dynamics, and Function using NetworkX**. (G. Varoquaux, T. Vaught, J. Millman, Eds.)Proceedings of the 7th Python in Science Conference. **Anais...**Pasadena, CA USA: 2008.

HAMELINCK, C. N.; VAN HOOIJDONK, G.; FAAIJ, A. P. C. Ethanol from lignocellulosic biomass: Techno-economic performance in short-, middle- and long-term. **Biomass and Bioenergy**, v. 28, n. 4, p. 384–410, 2005.

HANCOCK, T.; TAKIGAWA, I.; MAMITSUKA, H. Mining metabolic pathways through gene expression. **Bioinformatics**, v. 26, n. 17, p. 2128–2135, 2010.

HANSON, P. K. *Saccharomyces cerevisiae*: A Unicellular Model Genetic Organism of Enduring Importance. **Current Protocols in Essential Laboratory Techniques**, v. 16, n. 1, 2018.

HENRY, S. A.; KOHLWEIN, S. D.; CARMAN, G. M. Metabolism and regulation of glycerolipids in the yeast *Saccharomyces cerevisiae*. **Genetics**, v. 190, n. 2, 2012.

HERNÁNDEZ-ELVIRA, M.; SUNNERHAGEN, P. Post-transcriptional regulation during stress. **FEMS Yeast Research**, p. foac025, 13 maio 2022.

HIGGINS, V. J. et al. Yeast genome-wide expression analysis identifies a strong ergosterol and oxidative stress response during the initial stages of an industrial lager

fermentation. **Applied and Environmental Microbiology**, v. 69, n. 8, 2003.

HOANG, L. T. et al. Metabolomic, transcriptomic and genetic integrative analysis reveals important roles of adenosine diphosphate in haemostasis and platelet activation in non-small-cell lung cancer. **Molecular Oncology**, v. 13, n. 11, p. 2406–2421, 2019.

HOWE, K. L. et al. Ensembl Genomes 2020-enabling non-vertebrate genomic research. **Nucleic Acids Research**, 2020.

HUANG, C. W. et al. Hydrogen sulfide and its roles in *Saccharomyces cerevisiae* in a winemaking context. **FEMS Yeast Research**, v. 17, n. 6, 2017.

IMKER, H. J. 25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and Maintenance. **Frontiers in Research Metrics and Analytics**, v. 3, 2018.

ISHIKAWA, S. A. et al. A Fast Likelihood Method to Reconstruct and Visualize Ancestral Scenarios. **Molecular Biology and Evolution**, v. 36, n. 9, 2019.

JACKSON, L. P.; REED, S. I.; HAASE, S. B. Distinct Mechanisms Control the Stability of the Related S-Phase Cyclins Clb5 and Clb6. **Molecular and Cellular Biology**, v. 26, n. 6, 2006.

JACOBUS, A. P. et al. Comparative Genomics Supports That Brazilian Bioethanol *Saccharomyces cerevisiae* Comprise a Unified Group of Domesticated Strains Related to Cachaça Spirit Yeasts. **Frontiers in Microbiology**, v. 12, p. 687, 15 abr. 2021.

JAISWAL, S. et al. Systems Biology Approaches for Therapeutics Development Against COVID-19. **Frontiers in Cellular and Infection Microbiology**, v. 10, 2020.

JANG, Y.; LIM, Y.; KIM, K. *Saccharomyces cerevisiae* strain improvement using selection, mutation, and adaptation for the resistance to lignocellulose-derived fermentation inhibitor for ethanol production. **Journal of Microbiology and Biotechnology**, 2014.

JAYAKODY, L. N. et al. Engineering redox cofactor utilization for detoxification of glycolaldehyde, a key inhibitor of bioethanol production, in yeast *Saccharomyces cerevisiae*. **Applied Microbiology and Biotechnology**, v. 97, n. 14, 2013.

JIA, C. et al. Node Attribute-enhanced Community Detection in Complex Networks. **Scientific Reports**, 2017.

JIA, D. et al. *Yarrowia lipolytica* construction for heterologous synthesis of α -santalene and fermentation optimization. **Applied Microbiology and Biotechnology**, v. 103, n. 8, 2019.

JÖNSSON, L. J.; MARTÍN, C. Pretreatment of lignocellulose: Formation of inhibitory by-products and strategies for minimizing their effects. **Bioresource Technology**, 2016.

JUNG, H. et al. Twelve quick steps for genome assembly and annotation in the classroom. **PLoS Computational Biology**, v. 16, n. 11, 2020.

JUNQUEIRA, T. L. et al. Techno-economic analysis and climate change impacts of sugarcane biorefineries considering different time horizons. **Biotechnology for Biofuels**, v. 10, n. 1, p. 50, 14 mar. 2017.

KALVARI, I. et al. Rfam 14: Expanded coverage of metagenomic, viral and microRNA families. **Nucleic Acids Research**, v. 49, n. D1, 2021.

KAMMERER, D. et al. Polyphenol screening of pomace from red and white grape varieties (*Vitis vinifera* L.) by HPLC-DAD-MS/MS. **Journal of Agricultural and Food Chemistry**, v. 52, n. 14, 2004.

KANEHISA, M. et al. KEGG: New perspectives on genomes, pathways, diseases and drugs. **Nucleic Acids Research**, v. 45, n. D1, 2017.

KANG, K. et al. Linking genetic, metabolic, and phenotypic diversity among *Saccharomyces cerevisiae* strains using multi-omics associations. **GigaScience**, v. 8, n. 4, p. giz015, 1 abr. 2019.

KARP, P. D. et al. The BioCyc collection of microbial genomes and metabolic pathways. **Briefings in Bioinformatics**, v. 20, n. 4, 2018.

KATOH, K.; STANDLEY, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. **Molecular Biology and Evolution**, v. 30, n. 4, p. 772–780, 1 abr. 2013.

KHAN, Z. et al. The impact of technological innovation and public-private partnership investment on sustainable environment in China: Consumption-based carbon emissions analysis. **Sustainable Development**, v. 28, n. 5, 2020.

KHELLA, C. A. et al. Recent Advances in Integrative Multi-Omics Research in Breast and Ovarian Cancer. **Journal of Personalized Medicine**, v. 11, n. 2, p. 149, fev. 2021.

KILLEEN, D. J.; BOULTON, R.; KNOESEN, A. Advanced monitoring and control of redox potential in wine fermentation. **American Journal of Enology and Viticulture**, v. 69, n. 4, 2018.

KIM, I. S. et al. *Saccharomyces cerevisiae* KNU5377 Stress Response during High-Temperature Ethanol Fermentation. **Molecules and Cells**, v. 35, n. 3, p. 210, mar. 2013.

KIM, S. et al. Integrative clustering of multi-level omics data for disease subtype discovery using sequential double regularization. **Biostatistics**, v. 18, n. 1, 2017.

KITAGAKI, H.; TAKAGI, H. Mitochondrial metabolism and stress response of yeast: Applications in fermentation technologies. **Journal of Bioscience and Bioengineering**, v. 117, n. 4, 2014.

KLUYVER, T. et al. **Jupyter Notebooks – a publishing format for reproducible computational workflows**. (F. Loizides, B. Schmidt, Eds.) Positioning and Power in Academic Publishing: Players, Agents and Agendas. **Anais...IOS Press**, 2016.

KOHLER, M. Chapter 19 - Economic Assessment of Ethanol Production. Em: BASILE, A. et al. (Eds.). **Ethanol**. [s.l.] Elsevier, 2019. p. 505–521.

KONG, J. et al. GAAP: A Genome Assembly + Annotation Pipeline. **BioMed Research International**, v. 2019, 2019.

KORF, I. Gene finding in novel genomes. **BMC Bioinformatics**, 2004.

KURITA, O. Overexpression of peroxisomal malate dehydrogenase MDH3 gene enhances cell death on H₂O₂ stress in the ald5 mutant of *Saccharomyces cerevisiae*. **Current Microbiology**, v. 47, n. 3, 2003.

KWAK, S. et al. Enhanced isoprenoid production from xylose by engineered *Saccharomyces cerevisiae*. **Biotechnology and Bioengineering**, v. 114, n. 11, 2017.

LABIS. De Novo Sequence Assignment. 2021.

LANTZ, H. et al. Ten steps to get started in Genome Assembly and Annotation. **F1000Research**, v. 7, 2018.

LARSSON, S. et al. The generation of fermentation inhibitors during dilute acid hydrolysis of softwood. **Enzyme and Microbial Technology**, v. 24, n. 3–4, p. 151–159, 1999.

LARSSON, S. et al. Influence of lignocellulose-derived aromatic compounds on oxygen-limited growth and ethanolic fermentation by *Saccharomyces cerevisiae*. **Applied Biochemistry and Biotechnology**, v. 84–86, n. 1–9, p. 617–632, 2000.

LASCHOBBER, G. T. et al. Identification of evolutionarily conserved genetic regulators of cellular aging. **Aging Cell**, v. 9, n. 6, 2010.

LEGRAS, J. L. et al. Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. **Mol. Ecol.**, v. 16, n. 10, p. 2091–2102, maio 2007.

LI, K. et al. Extracellular redox potential regulation improves yeast tolerance to furfural. **Chemical Engineering Science**, v. 196, 2019.

LIMA, L. C. N. et al. Metabolic effects of p-coumaric acid in the perfused rat liver. **Journal of Biochemical and Molecular Toxicology**, 2006.

LIN, E.-B. Big data analysis in bioinformatics. **Journal of Biometrics & Biostatistics**, v. 08, n. 05, 2017.

LIN, Z.; ZHANG, Y.; WANG, J. Engineering of transcriptional regulators enhances microbial stress tolerance. **Biotechnology Advances**, v. 31, n. 6, p. 986–991, 2013.

LIU, C.-G.; QIN, J.-C.; LIN, Y.-H. Fermentation and Redox Potential. Em: **Fermentation Processes**. [s.l: s.n.].

LIU, J. F. et al. Outline of the biosynthesis and regulation of ergosterol in yeast. **World Journal of Microbiology and Biotechnology**, v. 35, n. 7, 2019a.

LIU, J.-J. et al. Investigating the role of the transcriptional regulator Ure2 on the metabolism of *Saccharomyces cerevisiae*: a multi-omics approach. **Applied Microbiology and Biotechnology**, v. 105, n. 12, p. 5103–5112, 1 jun. 2021.

LIU, Q. et al. Rewiring carbon metabolism in yeast for high level production of aromatic

chemicals. **Nature Communications**, v. 10, n. 1, p. 4976, 31 out. 2019b.

LIU, W. et al. From *Saccharomyces cerevisiae* to human: The important gene co-expression modules. **Biomedical Reports**, v. 7, n. 2, 2017.

LIU, Z. L.; MA, M. Pathway-based signature transcriptional profiles as tolerance phenotypes for the adapted industrial yeast *Saccharomyces cerevisiae* resistant to furfural and HMF. **Applied Microbiology and Biotechnology**, v. 104, n. 8, 2020.

LODI, T.; GUIARD, B. Complex transcriptional regulation of the *Saccharomyces cerevisiae* CYB2 gene encoding cytochrome b2: CYP1(HAP1) activator binds to the CYB2 upstream activation site UAS1-B2. **Molecular and Cellular Biology**, v. 11, n. 7, 1991.

LU, H. et al. Multi-omics integrative analysis with genome-scale metabolic model simulation reveals global cellular adaptation of *Aspergillus niger* under industrial enzyme production condition. **Scientific Reports**, v. 8, n. 1, p. 14404, 26 set. 2018.

LUCIO LOPES FERMENTEC, M.; CRISTINA DE LIMA PAULILLO, S. Tailored yeast strains for ethanol production: The process driven selection ALTFERM Project: fermentation with high concentration of ethanol for reduction of vinasses volume View project Yeast selection View project. 2015.

LUO, W. et al. GAGE: Generally applicable gene set enrichment for pathway analysis. **BMC Bioinformatics**, 2009.

LUO, W. et al. Pathview Web: User friendly pathway visualization and data integration. **Nucleic Acids Research**, v. 45, n. W1, p. W501–W508, 3 jul. 2017.

LUTTIK, M. A. H. et al. The *Saccharomyces cerevisiae* ICL2 gene encodes a mitochondrial 2-methylisocitrate lyase involved in propionyl-coenzyme A metabolism. **Journal of Bacteriology**, v. 182, n. 24, 2000.

MALECKI, M. et al. Mitochondrial respiration is required to provide amino acids during fermentative proliferation of fission yeast. **EMBO reports**, v. 21, n. 11, p. e50845, 2020.

MALINA, C.; LARSSON, C.; NIELSEN, J. Yeast mitochondria: an overview of mitochondrial biology and the potential of mitochondrial systems biology. **FEMS Yeast Research**, v. 18, n. 5, p. foy040, 1 ago. 2018.

MARSIT, S. et al. Evolutionary biology through the lens of budding yeast comparative genomics. **Nature Reviews Genetics**, v. 18, n. 10, 2017.

MARTÍNEZ-MATÍAS, N. et al. Toward the discovery of biological functions associated with the mechanosensor Mtl1p of *Saccharomyces cerevisiae* via integrative multi-OMICS analysis. **Scientific Reports**, v. 11, n. 1, p. 7411, 1 abr. 2021.

MAURYA, D. P.; SINGLA, A.; NEGI, S. An overview of key pretreatment processes for biological conversion of lignocellulosic biomass to bioethanol. **3 Biotech**, v. 5, n. 5, p. 597–609, 2015.

MCALOON, A.; TAYLOR, F.; YEE, W. Determining the Cost of Producing Ethanol from

Corn Starch and Lignocellulosic Feedstocks. p. 44, 2000.

MCLAREN, W. et al. The Ensembl Variant Effect Predictor. **Genome Biology**, 2016.

MEDINA, V. G. et al. Elimination of glycerol production in anaerobic cultures of a *Saccharomyces cerevisiae* strain engineered to use acetic acid as an electron acceptor. **Applied and Environmental Microbiology**, v. 76, n. 1, 2010.

MENDOZA-CÓZATL, D. et al. Sulfur assimilation and glutathione metabolism under cadmium stress in yeast, protists and plants. **FEMS Microbiology Reviews**, v. 29, n. 4, 2005.

MISTRY, J. et al. Pfam: The protein families database in 2021. **Nucleic Acids Research**, v. 49, n. D1, 2021.

MIZIK, T. Impacts of International Commodity Trade on Conventional Biofuels Production. **Sustainability**, v. 12, n. 7, p. 2626, jan. 2020.

MONTLLOR-ALBALATE, C. et al. Extra-mitochondrial Cu/Zn superoxide dismutase (Sod1) is dispensable for protection against oxidative stress but mediates peroxide signaling in *Saccharomyces cerevisiae*. **Redox Biology**, v. 21, 2019.

MORALES, J.; MENDOZA, L.; COTORAS, M. Alteration of oxidative phosphorylation as a possible mechanism of the antifungal action of p-coumaric acid against *Botrytis cinerea*. **Journal of Applied Microbiology**, v. 123, n. 4, p. 969–976, 2017.

MORENO, A. D. et al. Insoluble solids at high concentrations repress yeast's response against stress and increase intracellular ROS levels. **Scientific Reports**, 2019.

MORRISSETTE, V. A.; ROLFES, R. J. The intersection between stress responses and inositol pyrophosphates in *Saccharomyces cerevisiae*. **Current Genetics**, v. 66, n. 5, 2020.

MUDGE, J. M.; HARROW, J. The state of play in higher eukaryote gene annotation. **Nat Rev Genet**, v. 17, n. 12, p. 758–772, 1 dez. 2016.

MUELLER, L. P. et al. The effects of thermal and ethanolic stress in industrial strains of *Saccharomyces cerevisiae*. **Research, Society and Development**, v. 9, n. 10, p. e6819109091–e6819109091, 14 out. 2020.

MUKAI, N. et al. PAD1 and FDC1 are essential for the decarboxylation of phenylacrylic acids in *Saccharomyces cerevisiae*. **Journal of Bioscience and Bioengineering**, 2010.

N, N. et al. Analysing the Effect of Mutation on Protein Function and Discovering Potential Inhibitors of CDK4: Molecular Modelling and Dynamics Studies. **PLOS ONE**, v. 10, n. 8, p. e0133969, 7 ago. 2015.

NAGAMATSU, S. T. et al. Genome Assembly of a Highly Aldehyde-Resistant *Saccharomyces cerevisiae* SA1-Derived Industrial Strain. **Microbiology Resource Announcements**, v. 8, n. 13, p. e00071-19, 28 mar. 2019.

NGUYEN, L. T. et al. IQ-TREE: A fast and effective stochastic algorithm for estimating

maximum-likelihood phylogenies. **Molecular Biology and Evolution**, v. 32, n. 1, 2015.

NIJKAMP, J. F. et al. De novo sequencing, assembly and analysis of the genome of the laboratory strain *Saccharomyces cerevisiae* CEN.PK113-7D, a model for modern industrial biotechnology. **Microbial Cell Factories**, v. 11, n. 1, p. 1–17, 26 mar. 2012.

NILSSON, A. et al. Cofactor dependence in furan reduction by *Saccharomyces cerevisiae* in fermentation of acid-hydrolyzed lignocellulose. **Applied and Environmental Microbiology**, v. 71, n. 12, p. 7866–7871, 2005.

NOOKAEW, I. et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: A case study in *Saccharomyces cerevisiae*. **Nucleic Acids Research**, v. 40, n. 20, p. 10084–10097, 2012.

O'LEARY, N. A. et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. **Nucleic Acids Research**, v. 44, n. D1, 2016.

OLSSON, L.; NIELSEN, J. On-line and in situ monitoring of biomass in submerged cultivations. **Trends in Biotechnology**, v. 15, n. 12, 1997.

PAGE, A. J. et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. **Microbial genomics**, v. 2, n. 4, p. e000056, 1 abr. 2016.

PAL, S. et al. Big data in biology: The hope and present-day challenges in it. **Gene Reports**, v. 21, 2020.

PARAPOULI, M. et al. *Saccharomyces cerevisiae* and its industrial applications. **AIMS Microbiology**, v. 6, n. 1, 2020.

PERENTHALER, E. et al. Beyond the Exome: The Non-coding Genome and Enhancers in Neurodevelopmental Disorders and Malformations of Cortical Development. **Frontiers in Cellular Neuroscience**, v. 13, 2019.

PERTEA, G.; PERTEA, M. GFF Utilities: GffRead and GffCompare. **F1000Research**, v. 9, 2020.

PETERSEN, A. M. et al. Systematic cost evaluations of biological and thermochemical processes for ethanol production from biomass residues and industrial off-gases. **Energy Conversion and Management**, v. 243, p. 114398, 1 set. 2021.

PETROVA, V. Y. et al. Dual targeting of yeast catalase A to peroxisomes and mitochondria. **Biochemical Journal**, v. 380, n. 2, 2004.

POMPON, D. Yeast Physiology and Biotechnology. **Biofutur**, v. 1999, n. 189, 1999.

POSADINO, A. M. et al. Coumaric acid induces mitochondrial damage and oxidative-mediated cell death of human endothelial cells. **Cardiovascular Toxicology**, 2013.

PRADHAN, R. P. et al. The dynamics among entrepreneurship, innovation, and economic growth in the Eurozone countries. **Journal of Policy Modeling**, v. 42, n. 5,

2020.

QI, F. et al. Novel mutant strains of *Rhodospiridium toruloides* by plasma mutagenesis approach and their tolerance for inhibitors in lignocellulosic hydrolyzate. **Journal of Chemical Technology and Biotechnology**, 2014.

RAJ, T. et al. Recent advances in commercial biorefineries for lignocellulosic ethanol production: Current status, challenges and future perspectives. **Bioresource Technology**, v. 344, p. 126292, 1 jan. 2022.

RAMIŁOWSKI, J. A. et al. Functional annotation of human long noncoding RNAs via molecular phenotyping. **Genome Research**, v. 30, n. 7, 2020.

RAMIREZ-GAONA, M. et al. YMDB 2.0: A significantly expanded version of the yeast metabolome database. **Nucleic Acids Research**, v. 45, n. D1, 2017.

RAWLINGS, N. D. et al. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. **Nucleic Acids Research**, v. 46, n. D1, 2018.

REDDI, A. R.; CULOTTA, V. C. SOD1 integrates signals from oxygen and glucose to repress respiration. **Cell**, v. 152, n. 1–2, 2013.

REGENBERG, B. et al. Growth-rate regulated genes have profound impact on interpretation of transcriptome profiling in *Saccharomyces cerevisiae*. **Genome Biology**, v. 7, n. 11, p. R107, 14 nov. 2006.

REID, R. J. D. et al. Selective ploidy ablation, a high-throughput plasmid transfer protocol, identifies new genes affecting topoisomerase I-induced DNA damage. **Genome Research**, v. 21, n. 3, 2011.

REINOSO, F. A. M. et al. Fate of p-hydroxycinnamates and structural characteristics of residual hemicelluloses and lignin during alkaline-sulfite chemithermomechanical pretreatment of sugarcane bagasse. **Biotechnology for Biofuels**, 2018.

RICHARD, P.; VILJANEN, K.; PENTTILÄ, M. Overexpression of PAD1 and FDC1 results in significant cinnamic acid decarboxylase activity in *Saccharomyces cerevisiae*. **AMB Express**, 2015.

RIGDEN, D. J.; FERNÁNDEZ, X. M. The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection. **Nucleic Acids Research**, v. 47, n. D1, 2019.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics**, v. 26, n. 1, p. 139–140, 2010.

RONCORONI, M. et al. The yeast IRC7 gene encodes a β -lyase responsible for production of the varietal thiol 4-mercapto-4-methylpentan-2-one in wine. v. 28, n. 5, p. 926–935, ago. 2011.

RUDAN, M. et al. Normal mitochondrial function in *Saccharomyces cerevisiae* has become dependent on inefficient splicing. **eLife**, v. 7, 2018.

RUSSELL, J. B. Another explanation for the toxicity of fermentation acids at low pH: anion accumulation versus uncoupling. **Journal of Applied Bacteriology**, v. 73, n. 5, p. 363–370, 1992.

RUTHERFORD, J. C. et al. Nutrient and stress sensing in pathogenic yeasts. **Frontiers in Microbiology**, 2019.

SALAMEH, D. et al. Highlight on the problems generated by p-coumaric acid analysis in wine fermentations. **Food Chemistry**, 2008.

SALZBERG, S. L. Next-generation genome annotation: We still struggle to get it right. **Genome Biology**, v. 20, n. 1, 2019.

SAMBUSITI, C. et al. A comparison of different pre-treatments to increase methane production from two agricultural substrates. **Applied Energy**, v. 104, p. 62–70, 1 abr. 2013.

SANTIAGO, M.; GARDNER, R. C. The IRC7 gene encodes cysteine desulphydrase activity and confers on yeast the ability to grow on cysteine as a nitrogen source. **Yeast**, v. 32, n. 7, 2015.

SAYERS, E. W. et al. GenBank. **Nucleic Acids Research**, 2019.

SCALZITTI, N. et al. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. **BMC Genomics**, v. 21, n. 1, p. 1–20, 9 abr. 2020.

SCHREIBER, K. et al. Alternative splicing in next generation sequencing data of *saccharomyces cerevisiae*. **PLoS ONE**, v. 10, n. 10, 2015.

SHUMATE, A.; SALZBERG, S. L. Liftoff: Accurate mapping of gene annotations. **Bioinformatics**, v. 37, n. 12, 2021.

SIBIRNY, A. A. Yeast peroxisomes: Structure, functions and biotechnological opportunities. **FEMS Yeast Research**, 2016.

SIEVERS, F.; HIGGINS, D. G. The Clustal Omega Multiple Alignment Package. Em: **Methods in Molecular Biology**. [s.l: s.n.]. v. 2231.

SIMÃO, F. A. et al. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics**, 2015.

SINDHU, R.; BINOD, P.; PANDEY, A. Biological pretreatment of lignocellulosic biomass - An overview. **Bioresource Technology**, v. 199, p. 76–82, 2016.

SINGH, P.; SINGH, N. Role of Data Mining Techniques in Bioinformatics. **International Journal of Applied Research in Bioinformatics**, v. 11, n. 1, 2020.

SOSKINE, M.; TAWFIK, D. S. Mutational effects and the evolution of new protein functions. **Nature Reviews Genetics**, v. 11, n. 8, p. 572–582, ago. 2010.

STANKE, M. et al. AUGUSTUS: A b initio prediction of alternative transcripts. **Nucleic Acids Research**, 2006.

STUDENT. THE PROBABLE ERROR OF A MEAN. **Biometrika**, v. 6, n. 1, p. 1–25, 1

mar. 1908.

STURGES, H. A. The Choice of a Class Interval. **Journal of the American Statistical Association**, v. 21, n. 153, 1926.

SUBRAMANIAN, I. et al. Multi-omics Data Integration, Interpretation, and Its Application. **Bioinformatics and Biology Insights**, v. 14, 2020.

SURYA, B. et al. Economic growth, increasing productivity of smes, and open innovation. **Journal of Open Innovation: Technology, Market, and Complexity**, v. 7, n. 1, 2021.

SZKLARCZYK, D. et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. **Nucleic Acids Research**, 2019.

TAHERZADEH, M. J. et al. Physiological effects of 5-hydroxymethylfurfural on *Saccharomyces cerevisiae*. **Applied Microbiology and Biotechnology**, v. 53, n. 6, p. 701–708, 2000.

TAPIA CARPIO, L. G.; SIMONE DE SOUZA, F. Competition between Second-Generation Ethanol and Bioelectricity using the Residual Biomass of Sugarcane: Effects of Uncertainty on the Production Mix. **Molecules**, v. 24, n. 2, p. 369, 21 jan. 2019.

TARCA, A. L. et al. A novel signaling pathway impact analysis. **Bioinformatics**, v. 25, n. 1, p. 75–82, 2009.

TARKOWSKA, A. et al. Eleven quick tips to build a usable REST API for life sciences. **PLoS Computational Biology**, v. 14, n. 12, 2018.

TECHAPARIN, A.; THANONKEO, P.; KLANRIT, P. Gene expression profiles of the thermotolerant yeast *Saccharomyces cerevisiae* strain KKU-VN8 during high-temperature ethanol fermentation using sweet sorghum juice. **Biotechnology Letters**, v. 39, n. 10, p. 1521–1527, 2017.

TOKAR, T. et al. MirDIP 4.1 - Integrative database of human microRNA target predictions. **Nucleic Acids Research**, v. 46, n. D1, 2018.

TU, Q. et al. NCycDB: A curated integrative database for fast and accurate metagenomic profiling of nitrogen cycling genes. **Bioinformatics**, v. 35, n. 6, 2019.

TUMULURU, J. S. **Biomass Preprocessing and Pretreatments for Production of Biofuels: Mechanical, Chemical and Thermal Methods**. [s.l: s.n.].

VALL-LLAURA, N. et al. Redox control of yeast Sir2 activity is involved in acetic acid resistance and longevity. **Redox Biology**, v. 24, 2019.

VAN DER AUWERA, G. A. et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. **Current Protocols in Bioinformatics**, 2013.

VAN DER POL, E. et al. Analysis of by-product formation and sugar monomerization

in sugarcane bagasse pretreated at pilot plant scale: differences between autohydrolysis, alkaline and acid pretreatment. **Bioresource Technology**, v. 181, p. 114–123, abr. 2015.

VAN DER POL, E. C. et al. By-products resulting from lignocellulose pretreatment and their inhibitory effect on fermentations for (bio)chemicals and fuels. **Applied Microbiology and Biotechnology**, v. 98, n. 23, p. 9579–9593, dez. 2014.

VAN DONGEN, S.; ABREU-GOODGER, C. Using MCL to extract clusters from networks. **Methods in Molecular Biology**, 2012.

VAN ROSSUM G; DRAKE FL. Python 3 Reference Manual. **Scotts Valley, CA: CreateSpace**, 2019.

VERDUYN, C. et al. Effect of benzoic acid on metabolic fluxes in yeasts: A continuous-culture study on the regulation of respiration and alcoholic fermentation. **Yeast**, v. 8, n. 7, 1992.

VILLALBA, G. C.; MATTE, U. Fantastic databases and where to find them: Web applications for researchers in a rush. **Genetics and Molecular Biology**, v. 44, n. 2, 2021.

WALKER, R. S. K.; PRETORIUS, I. S. Applications of yeast synthetic biology geared towards the production of biopharmaceuticals. **Genes**, v. 9, n. 7, 1 jul. 2018.

WANG, R. et al. Integrative analyses of metabolome and genome-wide transcriptome reveal the regulatory network governing flavor formation in kiwifruit (*Actinidia chinensis*). **New Phytologist**, v. 233, n. 1, p. 373–389, 2022.

WOHLBACH, D. J. et al. Comparative Genomics of *Saccharomyces cerevisiae* Natural Isolates for Bioenergy Production. **Genome Biology and Evolution**, v. 6, n. 9, p. 2557–2566, 1 set. 2014.

YAHYA, F. A. et al. A brief overview to systems biology in toxicology: The journey from in to vivo, in-vitro and –omics. **Journal of King Saud University - Science**, v. 33, n. 1, 2021.

YANG, Y. et al. Progress and perspective on lignocellulosic hydrolysate inhibitor tolerance improvement in *Zymomonas mobilis*. **Bioresources and Bioprocessing**, 2018.

YUE, J. X. et al. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. **Nature Genetics** 2017 49:6, v. 49, n. 6, p. 913–924, 17 abr. 2017.

ZHANG, K. et al. Genomic reconstruction to improve bioethanol and ergosterol production of industrial yeast *Saccharomyces cerevisiae*. **Journal of Industrial Microbiology and Biotechnology**, v. 42, n. 2, p. 207–218, 1 fev. 2015.

ZHANG, K. et al. Genomic structural variation contributes to phenotypic change of industrial bioethanol yeast *Saccharomyces cerevisiae*. **FEMS Yeast Research**, v. 16, n. 2, p. 118, 1 mar. 2016.

ZHANG, K. et al. Genetic characterization and modification of a bioethanol-producing

yeast strain. **Applied Microbiology and Biotechnology**, v. 102, n. 5, p. 2213–2223, mar. 2018.

ZHANG, M. et al. Screening of thermosensitive autolytic mutant brewer's yeast and transcriptomic analysis of heat stress response. **Canadian Journal of Microbiology**, v. 66, n. 11, 2020.

ZHANG, M.; CASE, D. A.; PENG, J. W. Propagated Perturbations from a Peripheral Mutation Show Interactions Supporting WW Domain Thermostability. **Structure**, v. 26, n. 11, p. 1474- 1485.e5, 6 nov. 2018.

ZHANG, M. M. et al. Enhanced acetic acid stress tolerance and ethanol production in *Saccharomyces cerevisiae* by modulating expression of the de novo purine biosynthesis genes. **Biotechnology for Biofuels**, v. 12, n. 1, 2019.

ZHANG, T. et al. Increased heme synthesis in yeast induces a metabolic switch from fermentation to respiration even under conditions of glucose repression. **Journal of Biological Chemistry**, v. 292, n. 41, 2017.

ZHOU, Y. J. et al. Harnessing Yeast Peroxisomes for Biosynthesis of Fatty-Acid-Derived Biofuels and Chemicals with Relieved Side-Pathway Competition. **Journal of the American Chemical Society**, v. 138, n. 47, 2016.

ZITOMER, R. S.; LOWRY, C. V. Regulation of gene expression by oxygen in *Saccharomyces cerevisiae*. **Microbiological Reviews**, v. 56, n. 1, 1992.

ANNEX A - Bioethics and Biosafety Statement



Ministério do Meio Ambiente
CONSELHO DE GESTÃO DO PATRIMÔNIO GENÉTICO

SISTEMA NACIONAL DE GESTÃO DO PATRIMÔNIO GENÉTICO E DO CONHECIMENTO TRADICIONAL ASSOCIADO

Certidão
Cadastro nº A006ECE

Declaramos, nos termos do art. 41 do Decreto nº 8.772/2016, que o cadastro de acesso ao patrimônio genético ou conhecimento tradicional associado, abaixo identificado e resumido, no Sistema Nacional de Gestão do Patrimônio Genético e do Conhecimento Tradicional Associado foi submetido ao procedimento administrativo de verificação e não foi objeto de requerimentos admitidos de verificação de indícios de irregularidades ou, caso tenha sido, o requerimento de verificação não foi acatado pelo CGen.

Número do cadastro: **A006ECE**
 Usuário: **Universidade de São Paulo**
 CPF/CNPJ: **63.025.530/0001-04**
 Objeto do Acesso: **Patrimônio Genético**
 Finalidade do Acesso:
☒ **Pesquisa Científica** ☐ **Bioprospecção** ☐ **Desenvolvimento Tecnológico**

Espécie

Saccharomyces cerevisiae
Lactobacillus fermentum
Lactobacillus plantarum
Saccharomyces cerevisiae
Saccharomyces cerevisiae

Título da Atividade: **Fisiologia e Biotecnologia de Leveduras e Bactérias**

Equipe

Thiago Olitta Basso	Universidade de São Paulo
Bruno Labate Vale da Costa	Universidade Estadual de Campinas
Dielle Pierotti Procópio	Universidade de São Paulo
Rafael Ferraz Alves	Universidade Estadual de Campinas
Priscila Cola	Universidade de São Paulo
Felipe Eduardo Ciamponi	Universidade Estadual de Campinas
Andreas Karoly Gombert	Universidade Estadual de Campinas
Felipe Senne de Oliveira Lino	Novo Nordisk Foundation Center for Biosustainability
Luiz Carlos Basso	Universidade de São Paulo
Mario Lucio Lopes	Fermentec
Henrique Vianna de Amorim	Fermentec
Fernanda Sgarbosa Gomes Zanon	Amyris
Boris Juan Carlos Ugarte Stambuk	Universidade Federal de Santa Catarina
Aldo Tonso	Universidade de São Paulo

Morten Otto Alexander Sommer
Christopher Workman
Jette Thykær
Thomas Rasmussen
Gillian Eggleston
Marcelo Mendes Brandão

Novo Nordisk Foundation Center for Biosustainability
Technical University of Denmark
Technical University of Denmark
Novozymes
United States Department of Agriculture
Universidade Estadual de Campinas

Parceiras Nacionais

46.068.425/0001-33 / Universidade Estadual de Campinas

Resultados Obtidos

Divulgação de resultados em meios científicos ou de comunicação

Identificação do meio onde foi divulgado: Os resultados obtidos a partir dessa atividade

Data do Cadastro: 04/11/2018 13:04:09

Situação do Cadastro: Concluído



Conselho de Gestão do Patrimônio Genético
Situação cadastral conforme consulta ao SisGen em 20:37 de 02/08/2019.




SISTEMA NACIONAL DE GESTÃO
DO PATRIMÔNIO GENÉTICO
E DO CONHECIMENTO TRADICIONAL
ASSOCIADO - **SISGEN**

ANNEX B - Copyright Statement

Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada CARACTERIZAÇÃO DO PERFIL GENO-TRANSCRIPTÔMICO DE UMA LINHAGEM INDUSTRIAL DE *SACCHAROMYCES CEREVISIAE* EM RESPOSTA A ESTRESSE INDUZIDO POR ÁCIDO PARA-CUMÁRICO, não infringem os dispositivos da Lei n.º 9.610/98, nem o direito autoral de qualquer editora.


Campinas,

Documento assinado digitalmente
 FELIPE EDUARDO CIAMPONI
Data: 12/08/2022 10:18:06-0300
Verifique em <https://verificador.iti.br>

Assinatura : _____

Nome do(a) autor(a): Felipe Eduardo Ciamponi

RG n.º 50.320.283-6

Documento assinado digitalmente
 MARCELO MENDES BRANDAO
Data: 12/08/2022 16:28:39-0300
Verifique em <https://verificador.iti.br>

Assinatura : _____

Nome do(a) orientador(a): Marcelo Mendes Brandão

RG n.º 1.406.735