Universidade do Minho
Escola de Engenharia

Carolina Santiago Garrido Dias da Torre

Exploitation and annotation of *Torulaspora delbrueckii* genomes: comparison with biotechnological data

February 2021

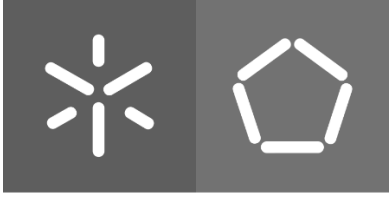
Exploitation and annotation of *Torulaspora delbrueckii* genomes: comparison with biotechnological data
Carolina Santiago Garrido Dias da Torre
UMINHO | 2021

Universidade do Minho
Escola de Engenharia

Carolina Santiago Garrido Dias da Torre

Exploitation and annotation of *Torulaspora delbrueckii* genomes: comparison with biotechnological data

February 2021



Universidade do Minho
Escola de Engenharia

Carolina Santiago Garrido Dias da Torre

Exploitation and annotation of *Torulaspora delbrueckii* genomes: comparison with biotechnological data

Master Degree in Bioinformatics

Dissertation supervised by

Ricardo Franco-Duarte, PhD

Pedro Soares, PhD

February 2021

ACKNOWLEDGEMENTS

First I want to thank to both my supervisors that guided and inspired me during this process and with whom I have learned so much.

I want to thank Simão for his help with the PCA. Teresa Rito who collaborated on the BLASTs shown in the first part of the work. Daniel who built the database of fungi used in taxonomic analysis.

Also a thanks to my supervisor's students Mafalda, Erica, Luísa and Ticiana, for their support.

I have to thank my family for always supporting me, and Tozé, for being my rock and being always there for me in the last years.

I also want to thank all my Bioinformatics master colleagues, for making this master's degree an even better experience, in special Joana, João and Cátia. Also want to thank all the teachers and Prof. Miguel Rocha for making this journey possible.

A special thanks to Juca, Jajão, Pedro and Filipe for being my discord company during those months of lookdown, helping me keep mentally sane and motivated.

This work was supported by "Contrato-Programa" UIDB/04050/2020 and project PTDC/BIA-MIC/32 059/2017 funded by national funds through the FCT, I.P. and by the ERDF through the COMPETE2020– Programa Operacional Competitividade e Internacionalização (POCI) and Sistema de Apoio Investigação Científica e Tecnológica (SAICT).

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

ABSTRACT

Nowadays, the most widely used yeast in wine, beer, and bread fermentations is *Saccharomyces cerevisiae*. However, in the past years, *Torulaspora delbrueckii* attracted interest due to its properties, from flavor and aroma-enhanced wine to the ability to be preserved longer in frozen dough. The main objective of this thesis was to explore *T. delbrueckii* genomes publicly available and the ones belonging to our project's collection, exploring their genomic information and establishing its relationship with their origins and biotechnological applications.

In the first phase, publicly available genomes of *T. delbrueckii* were explored, and their annotation was improved. EggNOG-mapper was used to perform functional annotation of the deduced *T. delbrueckii* coding genes, offering insights into its biological significance, and revealing 24 clusters of orthologous groups (COG), gathered in three main functional categories: information storage and processing (28% of the proteins), cellular processing and signaling (27%) and metabolism (23%). Small intra-species variability was found when considering functional annotation of the four *T. delbrueckii* available genomes. A comparative study was also conducted between *T. delbrueckii* genome and those from 386 fungal species, revealing high homology with species of *Zygotorulaspora* and *Zygosaccharomyces* genera, but also with *Lachancea* and *S. cerevisiae*. Lastly, the phylogenetic placement of *T. delbrueckii* was assessed using the core homologues found across 204 common protein sequences of 386 fungal species and strains.

In a second phase, the genome of fifty-four *T. delbrueckii* strains were sequenced and data was explored. The alignment, SNP statistics, annotation, among other steps, were attempted, for the first time, for those strains. PCA analysis was performed with those strains and the ones publicly available, to better understand the connection between the strains' technological groups.

The present work represents a successful effort to increase and improve the annotation of *T. delbrueckii*'s genome. Overall, this work provides a starting point to unravel the diversity of potential biotechnological applications of *T. delbrueckii*.

Keywords: Fermentation; NGS; *Saccharomyces cerevisiae*; *Torulaspora delbrueckii*; winemaking.

RESUMO

Hoje em dia, a levedura mais utilizada na fermentação de vinho, cerveja e pão é a *Saccharomyces cerevisiae*. No entanto, nos últimos anos a *Torulaspora delbrueckii* tem despertado interesse, devido às suas propriedades, desde o sabor e aroma do vinho até a capacidade de ser preservado por mais tempo em massa congelada. O principal objectivo desta tese foi explorar os genomas de *T. delbrueckii* publicamente disponíveis, assim como os que constituem a coleção do nosso projeto, analisando as suas informações genómicas e estabelecendo relação com suas origens e uso biotecnológico.

Na primeira fase, genomas publicamente disponíveis de *T. delbrueckii* foram explorados e sua anotação foi aprimorada. EggNOG-mapper foi usado para realizar a anotação funcional dos genes codificantes de *T. delbrueckii*, oferecendo uma perspetiva sobre seu significado biológico, o que revelou 24 grupos de grupos ortólogos (COG), reunidos em três categorias funcionais principais: armazenamento e processamento de informações (28% das proteínas), processamento e sinalização celular (27%) e metabolismo (23%). Pouca variabilidade intra-espécies foi encontrada quando se considerou a anotação funcional dos quatro genomas disponíveis de *T. delbrueckii*. Foi ainda realizado um estudo comparativo entre o genoma de *T. delbrueckii* e o de 386 espécies de fungos, o que revelou uma elevada homologia com espécies dos géneros Zygotorulaspora e Zygosaccharomyces, mas também com *Lachancea* e *S. cerevisiae*. Por último, foi avaliado o posicionamento filogenético da *T. delbrueckii* usando os homólogos encontrados em 204 sequências de proteínas comuns de 386 espécies e estirpes de fungos.

Na segunda fase, cinquenta e quatro estirpes de *T. delbrueckii* foram sequenciadas na Novogene e os dados recebidos foram explorados. Procedeu-se ao alinhamento, análise de SNP, anotação, entre outras análises, pela primeira vez, para essas estirpes. Com o objetivo de compreender a relação entre as estirpes estudadas, foi realizado um PCA.

O presente trabalho representa um esforço, bem-sucedido, para melhorar a anotação do genoma de *T. delbrueckii.* No geral, este estudo fornece um ponto de partida para desvendar as potenciais aplicações biotecnológicas da *T. delbrueckii*.

Palavras Chave: *Torulaspora delbrueckii*; *Saccharomyces cerevisiae*; NGS; fermentação; produção de vinho.

LIST OF ABBREVIATIONS AND ACRONYMS

BAM: Binary alignment map

BLAST: Basic Local Alignment Search Tool

CNVs: copy number variations

DNA: Deoxyribonucleic acid

IGV: Integrative Genomics Viewer

Indels: Insertions and Deletions

mtDNA: Mitochondrial DNA

NCBI: National Center for Biotechnology Information

NGS: Next generation sequencing

RNA: Ribonucleic acid

SAM: Sequence alignment map

SBS: Sequencing by synthesis

SNP: Single-nucleotide polymorphism

SNVs: single nucleotide variants

SVs: among other structural variants

VCF: Variant Call File

INDEX

FIGURE INDEX

TABLE INDEX

1 INTRODUCTION

1.1 Context

Torulaspora delbrueckii is a yeast, phylogenetically close to *Saccharomyces cerevisiae*. *T. delbrueckii* has gained attention due to its capacity to produce better wines and conserve longer in frozen dough. In this work, publicly available strains at NCBI were studied in order to enhance this species annotation and a place it phylogenetic.

In a second phase of this study, 54 strains isolated by our group were sequenced by Next Generation Sequencing (NGS) and the data obtained was analyzed. The main objective of this work was to expand the knowledge about this species and obtain information that latter will allow optimal use of *T. delbrueckii* in the industry.

1.2 Motivation

Torulaspora delbrueckii has been studied for its benefits when used in wine and beer fermentation, as well as in frozen dough. The main motivation for this thesis was to explore *T. delbrueckii* genomic information, possibly opening its use into those industries with enhanced benefits. To achieve that, not only publicly available strains were studied, but also 54 *T. delbrueckii* strains were isolated, their DNA was extracted, and then sequenced via NGS. The intensive study of those strains would expand information regarding diversity within the species. Moreover, it was important to comprehend *T. delbrueckii* adaptive evolution. Therefore, not only fermentation related information was explored, but the whole genome was subjected to an extensive study.

1.3 Objectives

In a first phase, this thesis had the objective of studying publicly available genomes of *T. delbrueckii*, and improve their annotation. A comparative study was conducted between *T. delbrueckii* genome and 386 other fungal species. At last, in this first phase, it was aimed to achieve the phylogenetic placement of *T. delbrueckii* regarding the remaining fungi strains that compose the database.

In a second phase the objective was to study and analyze the sequenced genomes of 54 *T. delbrueckii* isolated genomes. Firstly, an analysis of the received data was performed to confirm the quality of NGS

results. Thereafter, some statistical inferences were made and the annotation process took place, based on bioinformatics pipeline and tool's founded on homology search. To study the relation among all the available strains, Principal Component Analysis was applied. Geography, source of isolation and functional characteristics were studied to establish components that present a relation with those characteristics.

1.4 Document organization

Chapter 2 – State of the art

The aim of "state of the art" chapter is to introduce *T. delbrueckii*, the yeast species under study, and describe its importance in fields such as winemaking, beer brewing, bread fermentation, among others. Moreover, the methods currently available to perform the ambitioned analyses, are detailed and a perspective of the sequencing techniques evolution is summed up.

In the following chapters a division in two sections was made, based on the data used and the work´s objective:

A) Study of the publicly available genomes of *T. delbrueckii*.
B) Assess 54 *T. delbrueckii* newly sequenced strains, and study their genomes, aiming to understand their relationships.

Chapter 3 – Aims

A brief description of the aims to help contextualize the reader after the state of the art, to a better embark in the methods.

Chapter 4 – Materials and Methods

In this chapter the methodologies and the thinking behind the process are explained. At subchapter 4.1 *T. delbrueckii* genomes extracted from NCBI are studied. Functional annotation was achieved with EggNOG-mapper. A comparative study was conducted between *T. delbrueckii* genome and those from 386 fungal species. At last, the phylogenetic placement of *T. delbrueckii* was assessed using the core homologues found across 204 common protein sequences of 386 fungal species and strains. At subchapter 4.2, the received NGS data from Novogene® company is treated and analyzed. A comparative study between the strains is performed. At subchapter 4.3 the sequenced strains from our projects collection and all the *T. delbrueckii* strains available at NCBI are subject to a population analysis.

Chapter 5 – Results and discussion

In the "Results and discussion" chapter the information obtained is described and discussed. This section is once again divided in three sections in accordance with the Methods chapter. At subchapter 5.1, *T. delbrueckii* publicly available genomes are explored. At subchapter 5.2, the newly sequenced strains´ genomes are analyzed, their assemblies, SNPs, indels, among others are explored. The newly sequenced strains and all the *T. delbrueckii* genomes available at NCBI are explored and the results are compared with the strains biotechnological information available.

Chapter 6 – Conclusions and future work

In this chapter, the conclusions are taken and summarized and a reflection is made on the work done, allowing the establishment of necessary future work.

2 STATE OF THE ART

2.1 Yeasts and the human kind

Human kind and yeasts have been making history together for a long time. There are evidences that in prehistoric China, as early as the seventh millennium before Christ (B.C.), an alcoholic beverage was already being produced [1]. It is believed that the first industrial brewery was founded in the 17th century, by the Intendant Jean Talon, at a now known archeological site in the old part of Québec City [2]. A recent study conducted by Fijarczyk et al. [2], suggested that the study of isolates from this location could help reveal insights about human kind baking, brewing and winemaking history, therefore helping understand its migrations around the globe. This perspective highlights the straight connection between human kind and yeasts [2].

Figure 1. Palace of the Intendant of Quebec, 1759-1761. Hand-colored engraving by William Elliot (1727-1766) [3].

However, yeast history is not only made by their role in the alcoholic beverages production. In the more recent history, a yeast in particular, *Saccharomyces cerevisiae*, has played an important role in eukaryotic cell biology research since the first half of the 20th century. Identified by Louis Pasteur as the fermentation

agent [4], its genome was later the first eukaryotic genome sequenced [5]. This remarkable step was an impulse to the use of this yeast in developing fields like transcriptomics, metabolomics, proteomics, genome synthesis, genome editing, among others [4], [6]–[9].

2.2 *Torulaspora delbrueckii*

Torulaspora delbrueckii is a yeast known for its spheroidal to ellipsoidal cells. Asexual reproduction occurs by multilateral budding on a narrow base. True hyphae are never formed, however, pseudohyphae can be present [10]. *T. delbrueckii* has persistent asci, that might be conjugated, show conjugation between a cell and its bud or even between independent cells. In this species one may also find cells with tapered protuberances resembling conjugation tubes. This species presents asci with 1 to 4 spheroidal ascospores, that can be smooth or roughened [10].

Figure 2. *Torulaspora delbrueckii.* Extrated from [11].

T. delbrueckii is recognized for sugars fermentation, production of Coenzyme Q-6 and for having a negative reaction Diazonium blue B. Moreover, differences between *T. delbrueckii* strains regarding

fermentation capacities as well as carbon compounds assimilation are acknowledged, these being the reasons for the description of several taxa that later became synonyms [10]. As an example; *S. rosei*, *S. fermentati* and *S. vafer* are common synonyms of *T. delbrueckii* [12]. The anamorph of *T. delbrueckii* is *Candida colliculosa* [12].

T. delbrueckii is recognized for being highly resistant to numerous types of stress, among them salt and osmotic imbalance [13]. This yeast can be responsible for the spoilage of high sugars concentration foods (40-70%), which might be related with its capacity to tolerate low water activity conditions and osmotic stress. *T. delbrueckii* has also been considered as a great option to use in frozen dough products, due to its tolerance to freezing and freeze-thawing [10].

T. delbrueckii CBS 1146, the reference strain, is a wine isolate associated with an average fermentative performance speed and robustness [14], inferior to the commercially available strains, although with a higher production of glycerol. However, the strain COFT1 has been associated with increased fermentative performance and an ability to produce secondary metabolites, even higher than *S. cerevisiae* [15].

At the beginning of this study, a search in NCBI database for *T. delbrueckii* genomes, would reveal four assemblies. Of these, only the COFT1 strain had non-nuclear information. This was also the most recent of the four assemblies, having been published in March of 2018 and being labeled as complete. The CBS 1146 strain was the first to be deposited in December of 2011, and was last updated in April of 2018. This strain is the only one with available CDS from genomic FASTA available as well as being annotated. This strain currently stands as an assembly at chromosome level, on par with the NRRL Y-50541 strain, dated from June of 2015 and last modified in July of the same year. Finally, the SRCM101298 strain, added to the database in July of 2017, is considerably less explored, being assembled at the contig level. Later, in April of 2020, (27/04/2020) eleven *T. delbrueckii* strains assemblies were deposited in NCBI database, by the University College Dublin. All of these strains contained an assembly at scaffold level. With the addition of those eleven genes in April this year, the total number of available *T. delbrueckii* genomes at NCBI updated to fifteen.

2.3 Fermentation

Glycolysis, the process through which two molecules of pyruvate are obtained from one of glucose, is the principal pathway involved in ethanol fermentation. In mitochondrial respiration, in order to recycle the NADH formed in glycolysis, electrons pass to O_2 [16]. When in anaerobic conditions, ethanol is formed,

derived from the reduction of pyruvate, releasing CO_2. Yeasts cells biosynthesis require several energy dependent bioreactions; the ATP produced during glycolysis, is directed to those bioreactions, being yeast cell growth correlated with the process. It is essential that ATPs are being consumed in these reaction, otherwise its accumulation inside the cell inhibits phosphofructokinase, leading to the disruption of the glucose metabolism [17].

CO_2 and ethanol are not the only byproducts of ethanol fermentation, it is also observed, glycerol, higher alcohols, organic acids, lower flux of pyruvate and increased osmotic pressure [17]. Those aspects lead to alterations in parameters of fermented fresh foods such as taste, pH, texture, and preservation of food [16].

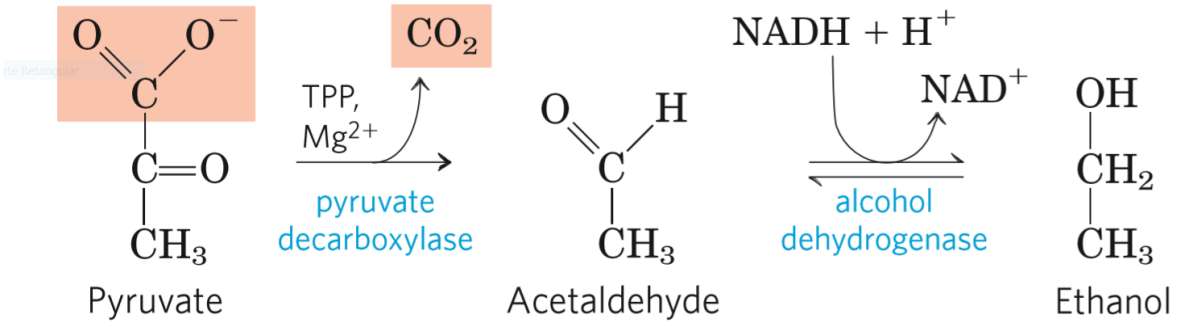


Figure 3. Conversion of pyruvate to ethanol and CO_2. Adapted from Nelson et al., (2017) [16].

Glycolysis allows conversion of glucose to pyruvate, as previously mentioned, and in a two-step process, pyruvate is converted to ethanol and CO_2 [16]. As demonstrated in Figure 3, in the first step, pyruvate decarboxylase catalyzes an irreversible decarboxylation of pyruvate. The second step consists of the reduction of acetaldehyde to ethanol mediated by alcohol dehydrogenase, relying on NADH reducing power [16]. The end products of ethanol fermentation are CO_2 and ethanol as shown in the overall equation below.

$$\text{Glucose} + 2\text{ADP} + 2\text{P}_\text{i} \longrightarrow 2\ \text{ethanol} + 2\text{CO}_2 + 2\text{ATP} + 2\text{H}_2\text{O}$$

All organisms that ferment glucose to ethanol have Pyruvate decarboxylase. This enzime has a role several aspects, for instance, the carbonation present in champagne, is due to the CO_2 resulting from pyruvate decarboxylation. In baking, it is responsible for the rising of dough [16].

Humans, like many other organisms that metabolize ethanol, possess alcohol dehydrogenase. In humans alcohol dehydrogenase is responsible for catalyzing the oxidation of ethanol in the liver, going in the contrary direction relative to the production of ethanol by fermentation [16].

2.3.1 Fermentation applications

Fermentation, is still nowadays, recognized by its role in production of alcoholic beverages, such as wine and beer. Brewing beer comprises numerous enzymatic processes. The essential ethanol fermentation is carried out by yeast glycolytic enzymes on carbohydrates in cereal grains. When added to the aerobic wort with sugars available to obtain their energy, yeasts grow and reproduce at a high rate. At this stage, abundant in O_2, the yeast oxidizes the pyruvate (obtained by glycolysis) producing CO_2 and H_2O via the citric acid cycle, without ethanol being formed. Once all the oxygen is consumed, the anaerobic metabolism starts, being now the sugars fermented into ethanol and CO_2. The fermentation process is mainly determined by parameters such as ethanol concentration, remaining sugar, and pH. Once those parameters dictate the end of fermentation, cells must be removed and the final processing can take place [16].

Large-scale production of alcoholic beverages led to the development of technology that is now being applied in other fields. One example is the production of ethanol, an environmentally friendly fuel. Ethanol is a good choice since its production is mainly performed using renewable resources that are relatively inexpensive [17]. These renewable resources must be rich in sucrose (beets or cane), starch (corn or wheat), or cellulose (straw, forest industry waste, or municipal solid waste). In a first step, monosaccharides must be obtained from raw materials, thereafter, in an industrial-scale fermenter, the monosaccharides are fed to a hardy strain of yeast. Besides ethanol, other side products can be produced that can fill other purposes, e.g., proteins being used for animal feed [16].

2.3.2 Yeasts and Wine Fermentation

S. cerevisiae has been the most commonly used yeast in the wine fermentation industry. The strains being currently used in wine fermentation industry are the result of a long process of selection trough fermentation history, being adapted to this purpose. Those yeasts have properties that allow a better performance regarding wine fermentation, and humans have even selected certain strains to certain wine types in order to enhance results [18]. *S. cerevisiae* wine strains have adapted in order to complete wine fermentation in conditions e.g., low pH (3.0–3.5), high alcohol content (up to 15% v/v), high sugar content (140–260 g/l), added sulphites (40–100 mg/l), and limiting amounts of nitrogen, lipids and vitamins [19].

Rossignol et al., in their experiments, were able to conclude that under wine fermentation conditions, subtelomeric genes are strongly regulated [19]. It is recognized that human fermentation have had impact on yeasts genome evolution [20]. And that, due to adaptation, subtelomeric genes were recently expanded in S. cerevisiae [21].

Moreover, yeasts are essential in wine production since they produce compounds of high relevance for wine flavor e.g., esters, higher alcohols, carbonyl compounds, volatile acids, volatile phenols and sulfur compounds [22]. The molecular bases behind the industrial strains are not totally understood, however polyploidy, aneuploidy and even rearranged chromosomes, can be observed in those strains, and in those alterations may rely the mechanisms that allow expression of certain genes or even copy number variation [23], [24].

2.4 *T. delbrueckii* in the food industry

T. delbrueckii made its way to wine industry as a complement to *S. cerevisiae* in the early phases of fermentation [25]. However, that is not the only reason why *T. delbrueckii* is recognized in food industry; from improved aromatic complexity and mouthfeel properties in wine [26] to the ability to be preserved longer in frozen dough, *T. delbrueckii* has received attention in the last years [27]. The first commercial *T. delbrueckii* option was a blend of it, *K. thermotolerans* and *S. cerevisiae*, that became available in 2003. Latter *T. delbrueckii* was released on its own [25], being now, at least one option of this yeast, available in the most dynamic yeast producers' catalogs [27].

Regarding the bakery industry, while *S. cerevisiae* loses approximately 80% of its viability when frozen at -20 °C in only 15 days, *T. delbrueckii* viability can be preserved for 4 months, making it a superior choice in this industry [27]. The benefits of *T. delbrueckii* use in cocoa bean's fermentation, comprehend quality enhancement of the end product, such as, alteration of the analytic profile and sensory perceptions of the chocolate, providing a different aroma profile [28]. With Durian fruit, has been reported that, not only it improves final quality of the product, due to aroma changes, but also, the fermentation process is completed [29]. The list of products that are benefited by the use of *T. delbrueckii* is extensive and includes lychee, mango, among others [27]. Due to the recent evidences, it became consensual, in the scientific community, that *T. delbrueckii* can enhance quality parameters. Improvements in aroma profiles are correlated with specific fruity esters, thiols, terpenes and low acetaldehyde production, while mouthfeel properties have been related to the releases of mannoproteins or polysaccharides that enhance sensory perception [27].

The main constraint regarding *T. delbrueckii* use in wine industry is it incapacity to complete the fermentation process by itself. Resulting in a less economically viable option than the mostly used process that employs *S. cerevisiae* [27]. However recent trends in wine industry have turned the attentions to aromatic and flavor profiles [25], conducting *T. delbrueckii* to the top of the non-Saccharomyces yeast [27].

Non-Saccharomyces yeasts contribution to flavor enhancement is mainly dependent on the production of metabolites, with special emphases to their concentration [25]. The selection of yeasts that develop in the medium is highly determined by must conditions such as the presence of SO_2, equimolar mixture of glucose and fructose, decreasing nutrients, high osmotic pressure, among others [25]. A proof of that is the reduction of the initial population of yeasts after clarification of white must, that comprehends centrifugation, enzyme treatments and cold settling [25][18].

Now, contrary to what was believed before, it is known that non-Saccharomyces yeasts do resist the beginning of alcoholic fermentation, not being the addition of SO_2 and the ethanol increasing, a death sentence for those yeasts [30], [31].

Jolly et all., considered three groups to classify non-Saccharomyces yeasts found in grape must during fermentation: Yeasts largely aerobic; Apiculate yeasts with low fermentative activity; and Yeasts with fermentative metabolism (*T. delbrueckii* belongs to this group). In the same article it is suggested a division of non-Saccharomyces in two groups, based on their flavor production; Neutral yeasts (low or no

flavor production); and Flavor-producing species (production of flavor compound, either desired or undesired) [25].

Wine content in esters is due to the equilibrium between two factors, the esters synthetizing enzymes and esterases. The first, such as ATF1, ATF2 and EHT1, are responsible for the esters synthesis, while the second is responsible for cleavage and sometimes ester bonds formation [22]. *S. cerevisiae* have a limited ability to liberate aromatic terpenols and other aglycones bound to saccharides. That is linked to β-Glucosidases, responsible for unleashing grape-derived aglycons, enhancing aroma and flavor in final wine [32].

In non-Saccharomyces yeasts, it has been observed different ways to transform odorless precursors in aroma compound. Moreover, is observed a high diversity in the aroma patterns formed [33]. As an example, co-inoculation with *T. delbrueckii* and *S. cerevisiae* registered higher concentrations of terpenols, C6 compounds and 2-phenylethanol, than mono-culture with *S. cerevisiae*. This appears to be the result of a cumulative effect between those yeasts metabolisms [34]. Usually when *T.delbrueckii* is used in sequential fermentation, lower levels of higher alcohols are observed. That is beneficial in wine industry, since it allows to increase the varietal character perception (which means it adds complexity to the final produced wine) [35].

Glycerol is an important metabolite, whose impact is dependent on wine style and grape variety. Its major contribution to wine final product is related with sweetness and smoothness [36]. However, when in high concentrations this metabolite can have negative effects in wine production since it increases acetic acid [27]. Non- Saccharomyces yeasts have been related with increased levels of glycerol, probably due to a more developed glycerol-3-phosphate dehydrogenase enzymatic activity, than alcohol dehydrogenase, leading to higher content of glycerol in fermentation [35]. It has been observed that in mixed cultures acetic acid and volatile compounds (like ethyl acetate) final concentrations are modulated, and polysaccharides production is increased [37].

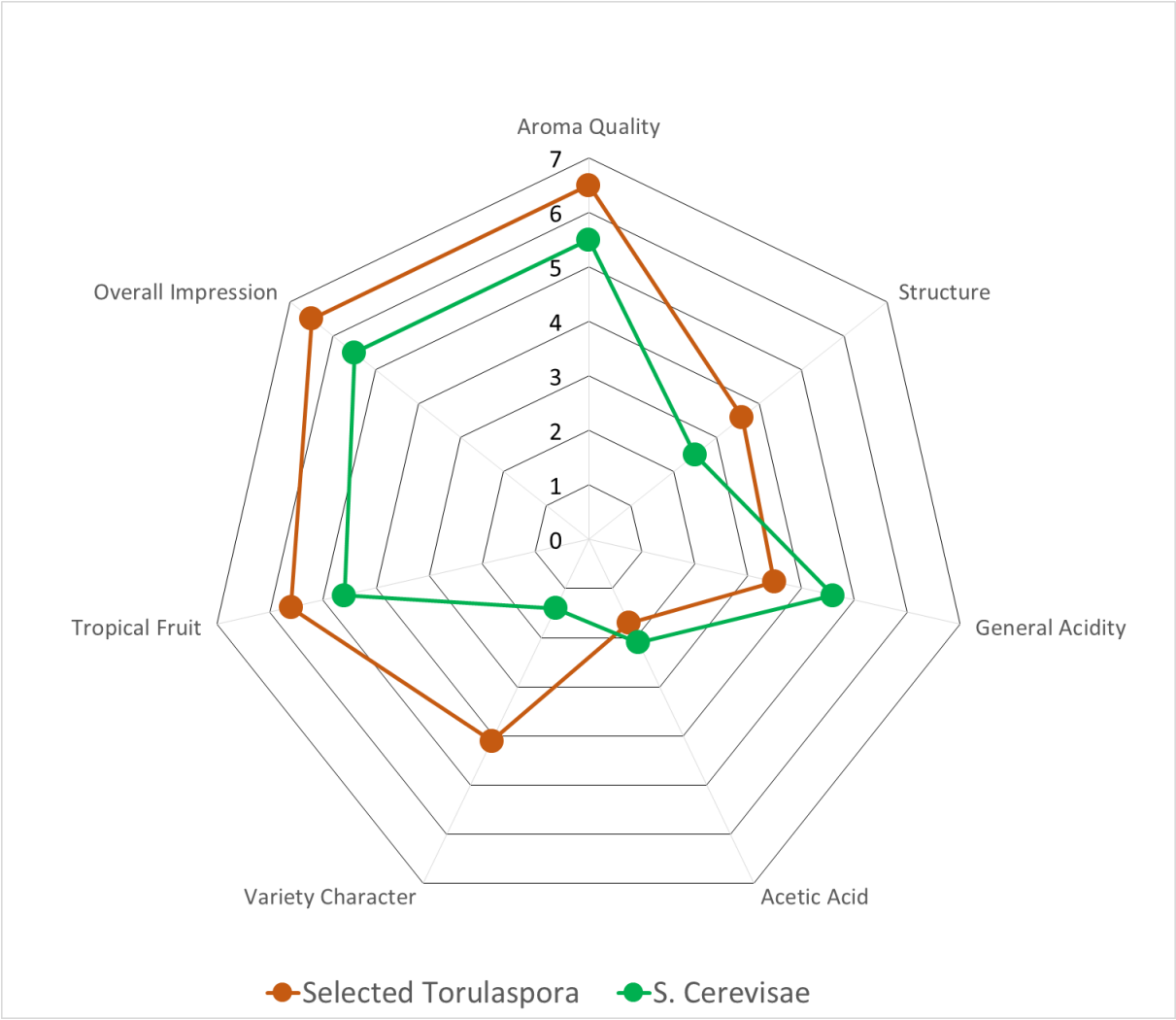


Figure 4. Summary of *T.delbrueckii* influences on sensory perception of fermented products compared to regular fermentations by *S. cerevisiae*. Extracted from Benito et al.,2018 [27].

Among the 680 compounds already found in wine, a great part of them have a role to play in the flavor and aromatization of wine, their effect is usually dictated by an optimized concentration window [38]. *T. delbrueckii* is recognized for modulating wine composition, aroma and flavor. Studies have been recorded that wines produced with coinoculation of *T. delbrueckii* and *S. cerevisiae*, show the enhancement of important parameters, for wine industry, such as fermentative esters, alcohols, lactones and fatty acids [39]. Nowadays, the more dynamic dry yeasts producers, have at least one strain of *T. delbrueckii* available [27].

Summarizing; it is recognized that *T. delbrueckii* brings advantages to wine such as low acetic acid and higher alcohols or high glycerol production, enhancing taste and aroma [15], [27], [40]. Not all about *T. delbrueckii* use is positive, and disadvantages include production of small amounts of 4-ethyl phenol,

increases in succinic acid concentrations and in precursors of some biogenic amines like histidine. But the most concerning disadvantage of *T. delbrueckii* is its inability to complete wine fermentation by itself. After that preview, becomes imperative to study this yeast, in order to obtain an optimized strain to apply in wine production, allowing to bring the advantages to the final wine.

2.5 Optimal strain

To obtain an optimal strain for wine production several aspects must be taken in account. First of all, in order to allow the production of low-ethanol wines and beer by itself, the optimal strain must be capable to ferment up to 9.5% (v/v) ethanol. Moreover, lower acetic acid concentrations, around 0.2 g/L, must be achieved. Once those principal points are achieved, the efforts should be focused in enhancing glycerol-pyruvic path to enable the production of softer wines, that require reduced ethanol levels. Then volatile profiles should be studied to obtain increases in terpene and thiolic contents and expand aromatic complexity. Moreover, attention should be given to ethyl phenols, succinic acid and biogenic amine precursors having in mind establish levels as lower as possible. Depending on the geographic area where the strain would be used, the malic acid degradation should be controlled to potentiate the strain. For use in red wine production, advantages could come from high pyruvic acid production, low anthocyanin adsorption, or well-developed hydrodynamic activity, therefore, for this purpose those points should also be monitored [27].

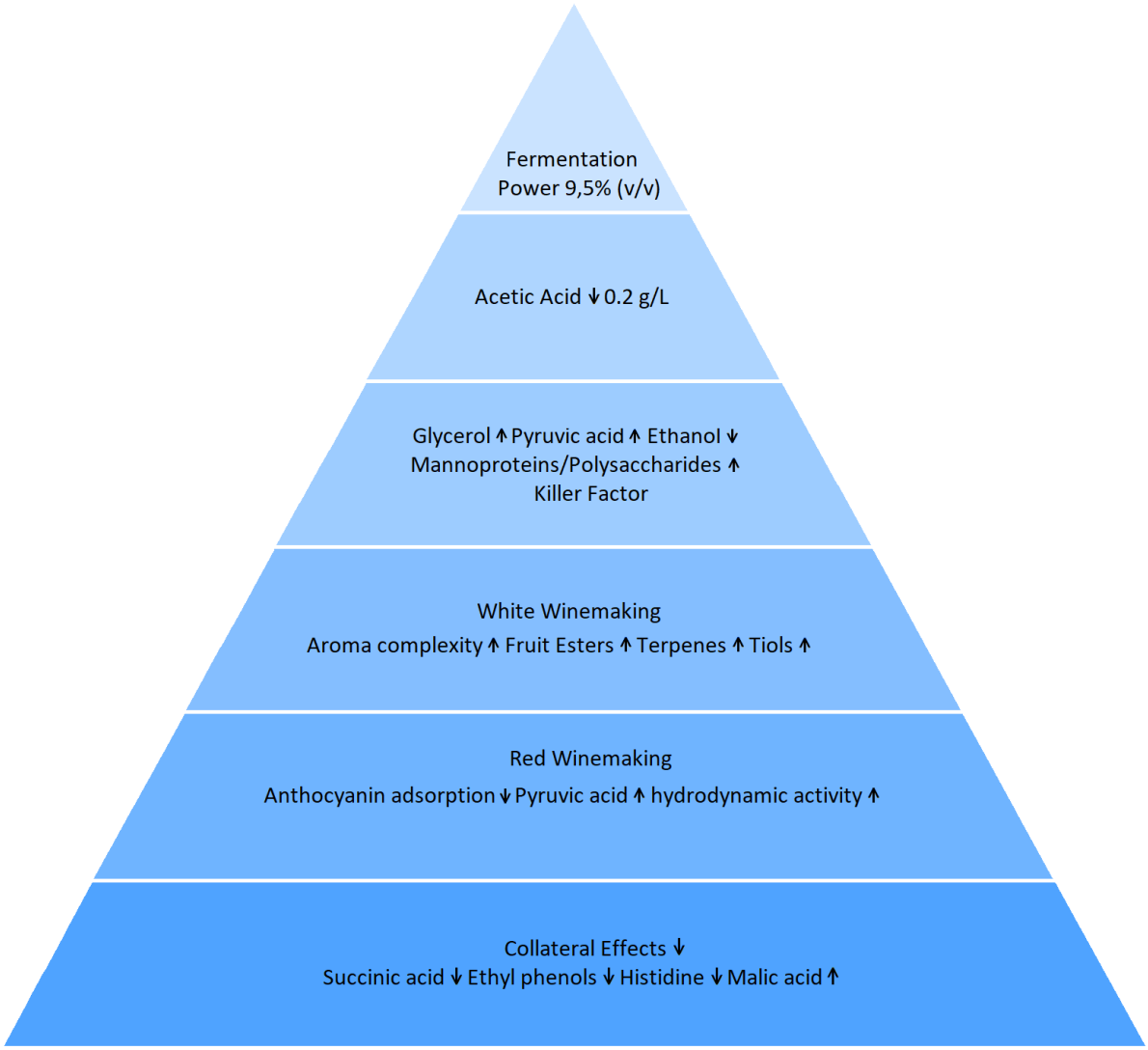


Figure 5. Proposed *T. delbrueckii* selection parameters. Extracted from Benito et al.,(2018) [27].

The introduction of an optimal strain in wine industry could be game changing. This improved strain could be either an optimized *T. delbrueckii* or a hybrid between *T. delbrueckii* and *S. cerevisiae*.

A conjugation of *S. cerevisiae* and *T. delbrueckii* may be an option to improve wine quality, if it is possible to combine the advantages of both strains. In the past a hybrid strain was created between the above mentioned species by Santos et al. (2012) [40] using protoplast fusion. The hybrid strain (F1-11) was characterized for the improved resistance to acetic acid and ethanol, as well as a fructose consumption similar to S. *cerevisiae*. F1-11 strain was able to perform the fermentation process by itself while improving flavor. Moreover, it was able to restart stuck fermentation [40].

2.6 Next-Generation sequencing

Sequencing is the process through which we can determine the DNA/RNA sequence of a genome. Over the years several techniques have been developed. In 1977, Sanger et al., publish an innovative DNA sequencing technique, considered more accurate and practical than the "plus and minus" method [41]. In 2005, bacterial vectors and Sanger sequencing were, still, the base for generating sequence information. At the time, sequencing a human genome was estimated in between $10 million and $25 million [42]. However, the introduction of NGS revolutionized the genomics field allowing the sequencing of hundreds of genomes in a short period for a much smaller price. The 454 technology was the first NGS platform created [43] in the year 2005, this technique is based on emulsion PCR and the measure of pyrophosphate (PPi), as visible light, while DNA is synthetized. Pyrosequencing was a remarkable step at the time for it accuracy, simplicity regarding the process, easiness to automatize, among others [44]. Those last years were marked by the investment in new sequencing technologies, conducting to an accelerated evolution of this area. In the table below are described some of those techniques.

Table 1. DNA sequencing technologies. Adapted from Morozova and Marra., 2008 [43].

Technology	Gen.	Approach	Read length	Bp per run	Company name
Automated Sanger sequencer	1st	Synthesis with dye terminators	Up to 900bp	96 kb	Applied Biosystems
454/Roche	2nd	Pyrosequencing by Synthesis	400-700 bp	80-129 Mb	Roche Applied Sciences
Solexa/Illumina		Sequencing by synthesis with reversible terminators	30-40 bp	8.5-600 Gb	Illumina, Inc.
ABI/SOLiD		Massively parallel sequencing by ligation	75+35 bp	90-180 Gb	Applied Biosystems
Chromium 10X		Tagged short reads	short	short	10x Genomics, Inc.
Ion semiconductor sequencing		Hydrogen detection from polymerization	35-400 bp	2-100 Gb	Ion Torrent Systems, Inc.
SMRT/PacBio sequencing	3rd	Real-time sequencing	~ 3000 kb	13 GB	Pacific Biosciences
Hi-C sequencing		Proximity-based ligation			Dovetail Genomics
MinION		Direct, real-time nanopore sequencing	Tens of Kb	Tens of GB	Oxford Nanopore Technologies

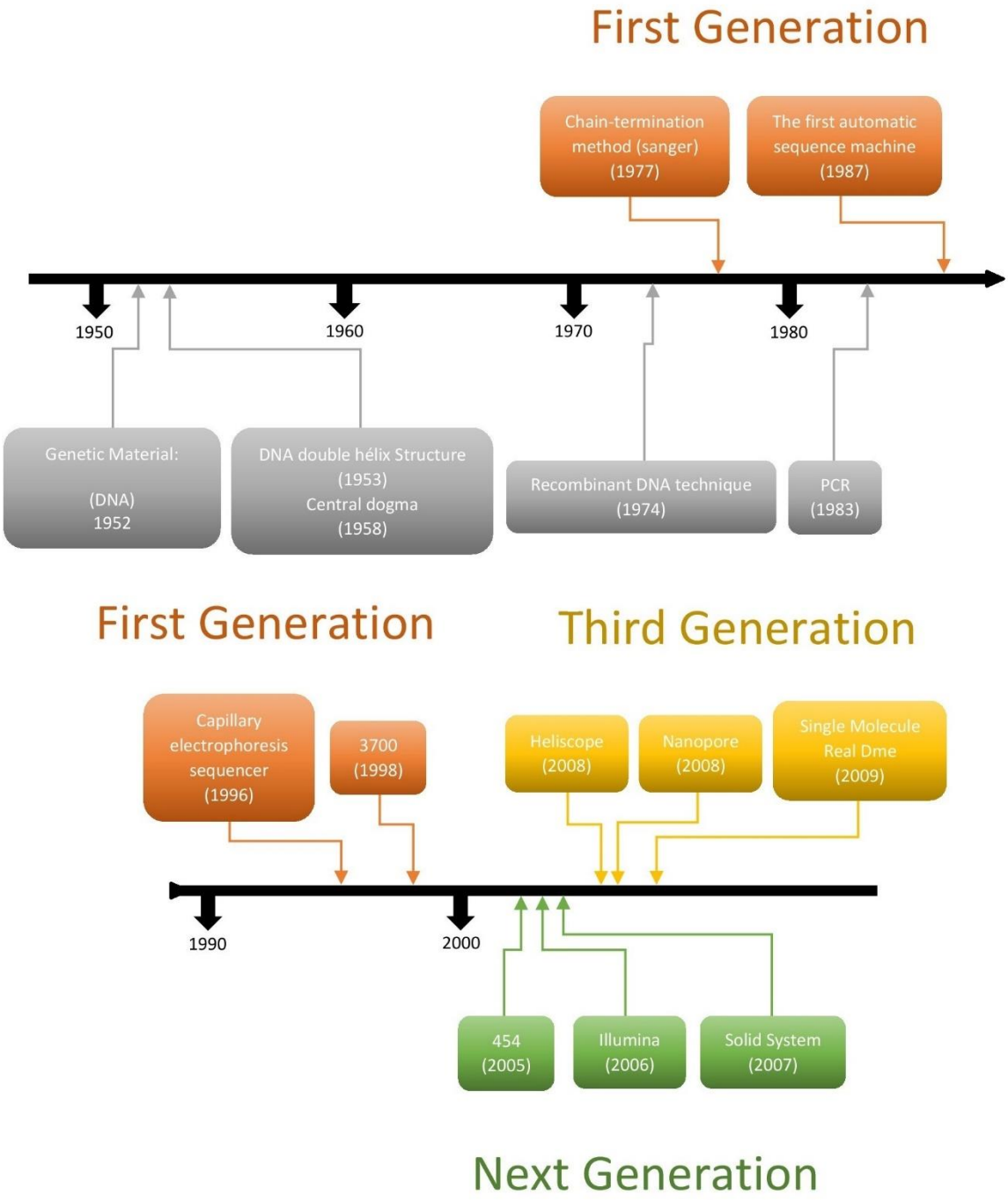


Figure 6. Sequencing Development Timeline adapted from a Novogene presentation.

Nowadays, the most used sequencing technology is Illumina. This technology can be divided in four major phases; library preparation, cluster generation, sequencing and data analysis. During library preparation are produced DNA fragments with adapters at both extremities. Inside each lane of the flow cell, are

oligos complementary to the adapters before mentioned. The attached sequences are amplified via PCR bridge amplification, originating several copies of those sequences (cluster). This produces a clonal amplification of the fragments since it occurs for millions of clusters at the same time. This process has been called "bridge amplification" since the DNA strands are obligated to arch in order to prime the next round of polymerization. Altered nucleotides (with fluorescence) are added to allow their identification, moreover, to do so, sequencing primers and DNA polymerase are also necessary. The primers hybridize to the sequences and DNA polymerase extends them with the fluorescent nucleotides. Later, a laser leads the clusters to emit light signals, those signals are then analyzed with bioinformatics tools, allowing to identify the nucleotide sequence [45].

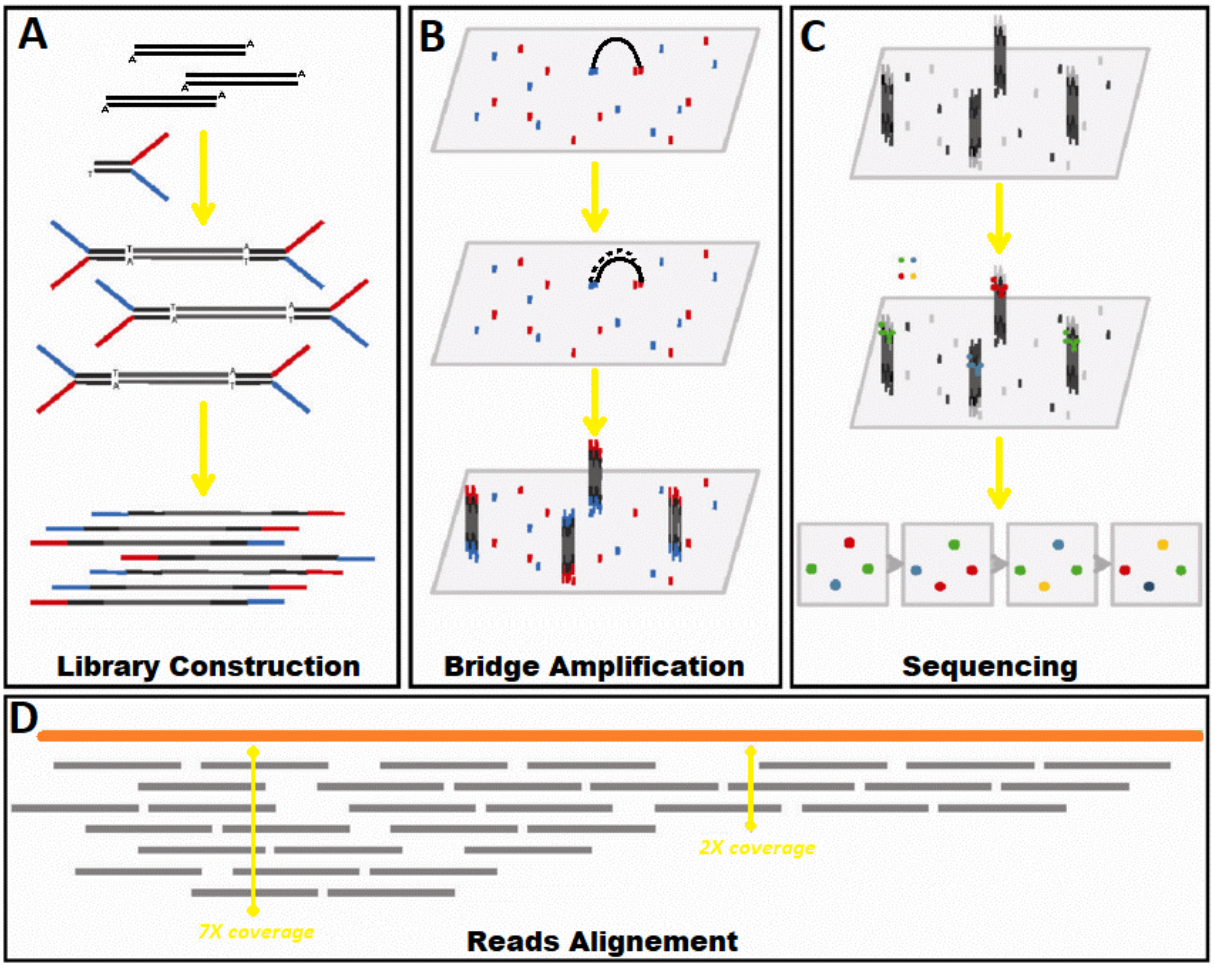


Figure 7. Illumina sequencing technology. Adapted from Brind'Amour (2010) [46]

There are several sequencing platforms for illumina: iSeq, MiSeq, MiniSeq, NextSeq, HiSeq and NovaSeq. Each of these platforms are optimized for a purpose. For example, NextSeq series are adapted for small genomes, targeted gene sequences and transcriptomic sequencing. HiSeq platforms are directed for whole-transcriptome sequencing and exome sequencing. When the aim is to achieve whole genome sequencing (WGS) NovaSeq is the appropriate platform [47], [48].

Limitations pointed in Sanger sequencing, e.g., low-resolution genotyping of mtDNA markers, inability to analyze multiple genetic polymorphisms in a single reaction using a single workflow, were overcome with NGS [49]. Nowadays, 90% of the world's sequencing data is produced by Illumina, sequencing by synthesis (SBS) [50].

2.6.1 How to deal with NGS data

As an illustration of the evolution in the sequencing field, before NGS, the Human Genome Project, the first human genome assembly, was a 13 years' project, that involved 3.4 billion USD, and the collaboration of hundreds of international labs. Since NGS was introduced in 2007, the price of sequencing a genome started to go down, very fast. Subsequently, in 2014, with the introduction of HiSeq X platform the price of sequencing a human genome was established in 1,000 USD, and, in only 3days, 16 human genomes could be sequenced [51]. This was without any doubt a big turn in bioinformatics, leading to an increasing in the amount of data generated in this science.

Nowadays, communal NGS applications comprise ChIP-seq, methyl-seq, RNA-seq and DNA-seq. The latter, can be applied to particular regions, whole exome sequencing (WES) or whole genome sequencing (WGS). The objective, when performing DNA sequencing, usually is to unravel genomic variations as: insertions or deletions (indels); copy number variations (CNVs); single nucleotide variants (SNVs); among other structural variants (SVs) [52].

After acquiring NGS data, some steps are essential to achieve relevant information. Analytical steps i.e., raw data quality analysis, pre-processing, reads alignment, post-processing, variant analysis are essential [52]. Figure 9 outlines the process, in a very general sense.

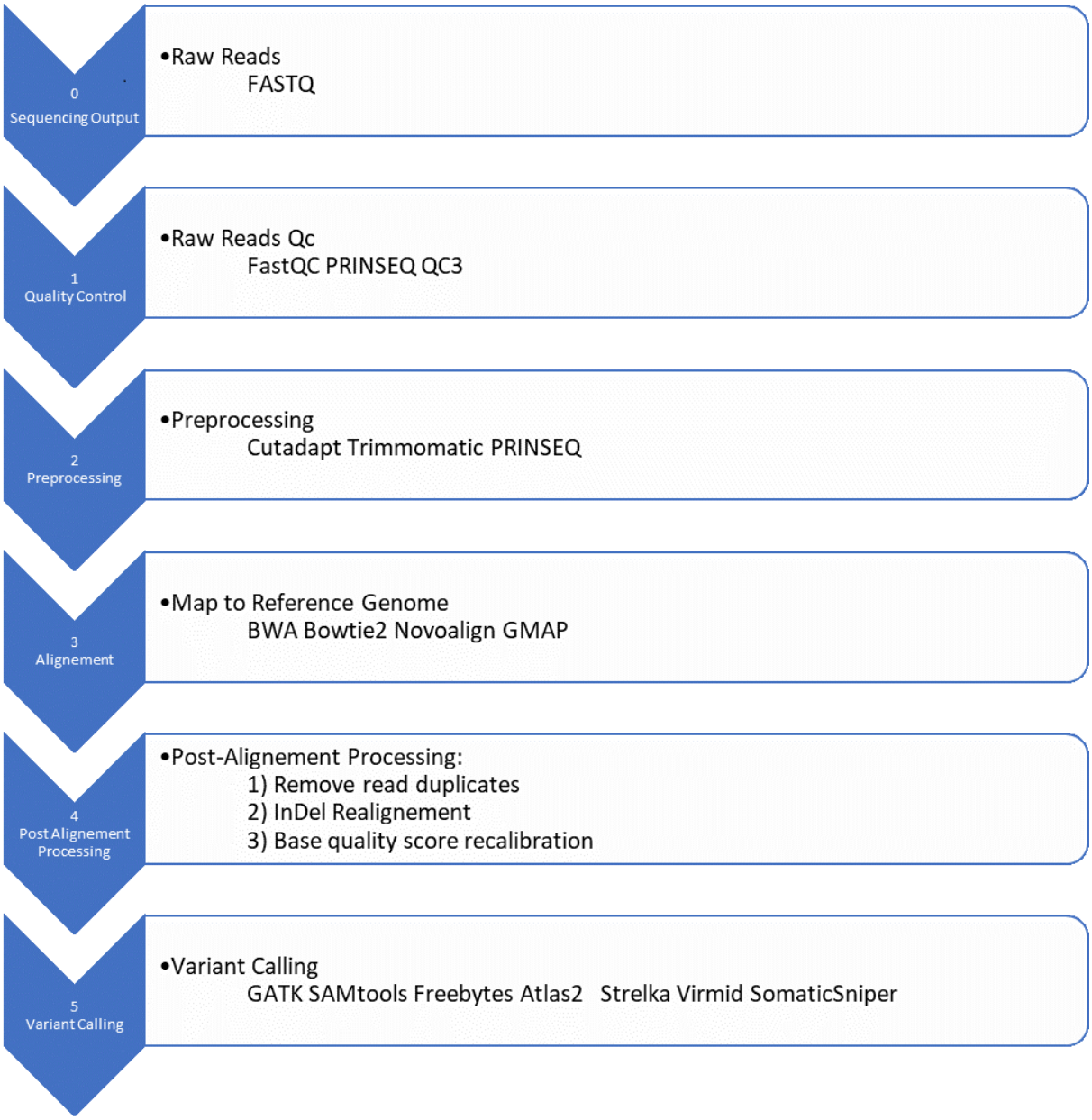


Figure 8. Initial steps in NGS data analysis. Adapted from Bao et al. (2014) [52].

2.6.2 FastQC

FastQC is "a quality control tool for high throughput sequence data". It requires Picard SAM/BAM libraries and a suitable Java runtime environment [53].

Formats accepted by FastQC include FastQ: GZip compressed FastQ; sequence alignment map (SAM), binary alignment map (BAM) [54]. The obtained analyses comprise; Basic Statistics; Per Base Sequence Quality; Per Sequence Quality Scores; Per Base Sequence Content; Per Base GC Content; Per Sequence

GC Content; Per Base N Content; Sequence Length Distribution; Duplicate Sequences; Overrepresented Sequences; Overrepresented Kmers [54].

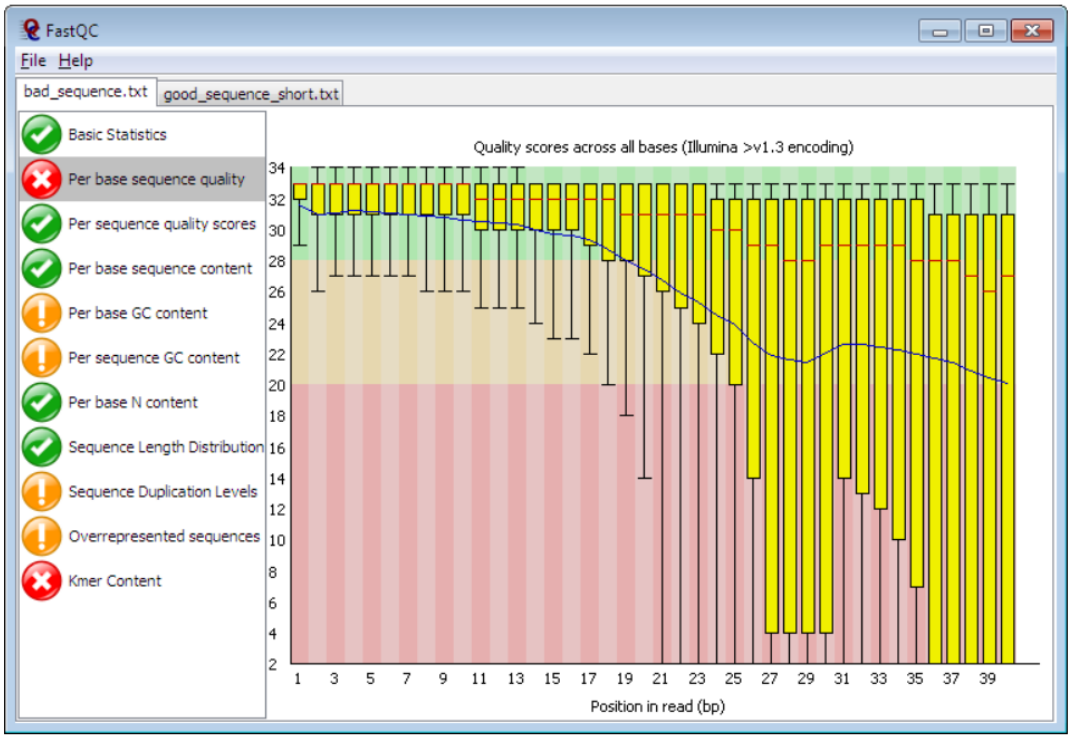


Figure 9. Per base sequence quality representation of FastQC output. This sample fails in this module due to having a significant amount of base positions with low quality scores. The low scores of this sample at the end is normal, due to the degradation of most sequencing platforms towards the end of the read. Image extracted from Babraham Bioinformatics [53].

Figure 9 exemplifies the output from FastQC. These analyses must be interpreted to optimize downstream analyses.

2.6.3 Genome assembly

Reads obtained from NGS sequencing must be assembled into a full genome. There are two types of Assembly; *de novo* assembly and with a reference genome. *De novo* assembly is a highly complex process in which the several fragments obtained by sequencing have to be assembled as a result of their shared regions [55] without any previous knowledge of the genome's composition. When working with a species that already have a reference genome (a genome that was already sequenced and assembled), the alignment can be guided by that reference genome [56].

Nowadays several tools are available to perform assembly such as: Velvet [57]; ABySS [58]; BWA [59]; Bowtie [60]; SPAdes [61]; SOAP2 [62]; among others. BWA is grounded on backward search with Burrows–Wheeler Transform (BWT), allowing gaps and mismatches. Moreover, BWA outputs the alignment in SAM format, allowing to use SAMtools directly. BWA is linked to human genome assembly, since it allows to map short sequences against a large reference genome [63].

In 2010, platforms were already producing longer reads and as it seemed to be the tendency, therefore, a new tool was created, more directed to this type of data, the Burrows-Wheeler Aligner's Smith-Waterman Alignment (BWA-SW) [59].

Nowadays, BWA supports three algorithms; BWA-backtrack, BWA-SW and BWA-MEM. BWA-backtrack was designed having in mind Illumina sequence reads up to 100bp. BWA-SW and BWA-MEM were developed for sequences in a range from 70bp to 1Mbp and split alignment. The most recent, BWA-MEM, is recommended for high-quality queries and for 70-100bp Illumina reads, since it is more accurate and fast [64].

2.6.4 Samtools

Samtools is a group of programs, that allow to work with high-throughput sequencing data. It is a collection of three repositories; Samtools, BCFtools and HTSlib. Samtools allows to Read, write, edit, index and viewing in three formats (SAM, BAM and CRAM). BCFtools permits to read and write (BCF2, VCF and gVCF) and also, call, filter and summarize SNP and short indel sequence variants. HTSlib ia a C library directed to reading and writing high-throughput sequencing data [65].

2.6.5 NGS Data Visualization

Integrative Genomics Viewer (IGV) [66], is a high-performance viewer that supports next-generation sequencing data. It is available for free download and it is user-friendly. The main goal of IGV is to allow visualization and exploration of data sets for researchers. With that in mind, IGV offers high-performance data visualization and exploration on standard desktop systems, moreover, it allows loading of local and remote data [66].

There are others software's that allow NGS data visualization: Tablet [67], BamView [68], Savant [69] and Artemis [70]. IGV outstands from them for two reasons. First, it allows to visualize data in multiple

genomic regions at the same time, in adjacent panels. Second, it supports different data types beyond NGS, such as array-based platforms, like expression or copy-number arrays that can be integrated, and when combined with metadata may allow grouping, sorting and filtering of the information [66].

IGV allows NGS data visualization, being an important ally to visualize aligned reads, allowing to confirm and interpret variant calls. Since 2009 it allows NGS data visualization, and have been developed to offer better tools to inspect, interpret and validate genomic data [71].

2.6.6 Genome annotation

Genome annotation is the assigning of meaningful biological information to a genome sequence. This process relies on the study of the genome sequence's structure and composition as well as the knowledge of reference cases, from closely related species. Usually the efforts are directed to the precise identification of protein coding genes [72].

The complexity of the annotation process and the amount of data needed to perform it, depends greatly on genome properties such as size, repeats, heterozygosity, ploidy level and content in GC [72]. Hence, preceding the annotation process, is necessary to collect and analyze data regarding the previously mentioned aspects. *T. delbrueckii* genome information available in NCBI stands for 8 chromosomes, 9.52Mb length, 4970 proteins and a 41.9% GC content. The species most used in wine production, *S. cerevisiae* has a genome composed by 16 chromosomes, a length of approximately 11.86Mb, 5404 proteins and a GC content of 38.3% [5]. Yeast species genomes tend to be compact and with few introns, nevertheless, even in cases of deep phylogenetic distances, they retain extensive synteny [73], meaning the conservation of the relative physical location of genes along a chromosome. The resources available and the results expected determine the approach followed to perform genome annotation. Being the full genome annotation, mostly, a very exploratory analysis combining many approaches. Mudge et al. (2016) [74], affirms that computational annotation is based on three methods: alignment, comparative annotation and ab initio annotation. The alignment stands for the alignment of transcript evidence. The comparative annotation, relies on the construction of models of the genome in progress based on closed species genomes. The annotation with use of algorithms, such as AUGUSTUS and GENSCAN, is called "Ab initio annotation", this approach allows to obtain models centered on a priori knowledge of their likely sequence [74].

Angel et al. (2018) [72], reports only two approaches to genome annotation: intrinsic and extrinsic. The intrinsic approach relies on the elaboration of statistical models, training and optimizations of the software. At this point, well annotated genes, needed in order to build models and train software, are crucial. Since every genome is different, the software and the models shall be specific for each case under study. The extrinsic approach is more generally applicable. The base is to explore an ample collection of described polypeptide sequences available in databases such as NCBI, RefSeq or UniProt, in order to reach gene prediction [72]. Eukaryotic genome annotation is usually a combination of an *ab initio* approach and extrinsic information [72], [74]. The Yeast Genome Annotation Pipeline (YGAP) is a practical solution to perform *ab initio* annotation of yeast genomes. When comparing YGAP and AUGUSTUS, Proux-Wéra et al. (2012) [73], concluded that YGAP had a better performance. Proux-Wéra et al. (2012) [73] describe YGAP as "an automated system designed specifically for new yeast genome sequences lacking transcriptome data". Yeast Gene Order Browser (YGOB) database has stored information concerning homology and synteny of yeast species. YGAP achieves automatic *de novo* annotation, thru the information deposited in YGOB. This relies on the premise that orthologous genes probably have similar intron/exon structures and that from the data present in YGOB, genes in specific genomic regions are predictable. Moreover, YGAP recognizes errors in frameshift sequencing and suggests corrections, detects transposable elements and tRNA genes and searches intelligently for introns [73].

EggNOG-mapper tool was developed to perform "functional annotation of large sets of sequences based on fast orthology assignments using precomputed clusters and phylogenies from the eggNOG database" [75]. This tool was validated by benchmarking the Gene Ontology (GO) predictions against BLAST and InterProScan, two well recognized homology-based approaches [75].

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a reference knowledge base that allows to obtain Orthology annotations. KEGG stablishes the connection between genomes and pathways with "a collection of manually defined ortholog groups identified by K numbers" [76].

Angel et al. (2018) [72] instructs that the methods should be computationally repeatable and reproducible, allowing investigation, reanalysis and re-annotation, conducting to a successful annotation project.

2.6.7 Phylogenetic analysis

Phylogenetics consists in the study of the relationships between individuals inferring their evolutionary history. The results are usually presented as a phylogenetic tree [77]. Several biological information can be used as molecular markers to construct a phylogenetic tree, being one of them DNA sequences, being possible to use both coding and non-coding regions. After selecting the biological information, it is necessary to proceed to the selection of homologous sequences. Then, those sequences must be subjected to a multiple sequence alignment (MSA), in order to ensure that every nucleotide in every sequence is compared with the homologous in the others sequences [78].

The resulting alignment will be the foundation to calculate the divergences and infer the samples evolutionary relationships. Therefore, this is an important step, in this sense, several algorithms have been developed in the last years, using different strategies. Nowadays, the alignment methods can be classified in three groups based on the approach they follow; progressive approach; consistency-based methods and the statistical or evolution-based methods. The progressive approach is the most common, and include Muscle52, MAFFT54 and Clustal53. The consistency-based methods, include ProbCons56, T-Coffee55 and some versions of MAFFT54. The statistical or evolution-based methods are the more expensive in a computationally perspective and comprehend StatAlign59 and Bali-Phy58 [78].

The programs used to generate the phylogenetic trees are mainly based in one of two strategies; A) distance-based matrix methods; or B) the character based methods [79]. The distance-based matrix methods calculate the genetic distances between every sequence. Then the previously calculated distances are computed on the proportion of different sites between the sequences originating a distance matrix. This matrix is then subjected to agglomerative clustering algorithms generating a phylogenetic tree. While, the character-based methods focus on the information at each homologous site of the sequences, involving the generation of several trees, that are then selected, to, finally obtain the best tree. Character-based methods attempt to reconstruct hierarchically the occurrence of mutation events according to a timeline, and in that sense it is a better representation of evolution. Different character-based have different principles. Parsimony algorithms are usually less-computational demanding and mostly used for short evolutionary distances, while maximum likelihood and Bayesian inference are more computational demanding, complex and are more accurate for deep phylogenies [80].

3 AIMS

This project aims can be divided in two objectives/sections;

A) To analyze *T. delbrueckii* publicly available genomes.
B) To study 54 *T. delbrueckii* strains newly sequenced, concerning their genomes and relationships.

With those analyses we aim to improve *T. delbrueckii* genome annotation, understand its phylogenetic placement in comparison with other fungi species, and understand the correlation between the strains origin and biotechnological use and their genomes. In detail, the obtained knowledge will allow to associate the strains' phenotypic differences with differences detected in their genome. Our final objective is to globally conclude about future strain enhancements, that could lead to an improvement of *T. delbrueckii* fermentation performance.

4 MATERIALS AND METHODS

4.1 Analysis of NCBI strains

4.1.1 Genome annotation

With the purpose of performing genome annotation, the four *T. delbrueckii* genomes available in NCBI database were considered: CBS 1146 (Accession number GCA_000243375.1), COFT1 (GCA_003013175.1), NRRL Y-50541 (GCA_001029055.1) and SRCM101298 (GCA_002214845.1). The four genomes were downloaded from NCBI and submitted to the Yeast Genome Annotation Pipeline (YGAP) [73], in order to establish potential coding regions in each of the *T. delbrueckii*'s chromosomes. Relevant information was extracted, in particular start and end positions of coding regions, strain orientation and known homologs in the reference genome of S. cerevisiae (strain S288c). The potential coding regions reported by YGAP were extracted from the complete *T. delbrueckii* genome into a FASTA file.

Proteins identified by YGAP as having no homology with S. cerevisiae were scrutinized by BLAST (Basic Local Alignment Search Tool), and results were combined assessing the number of taxonomic correspondences (top-hits) for each protein.

Functional genomic annotation was performed with eggNOG-mapper [75], considering proteins predicted by YGAP, and results were described considering Gene Ontology (GO) terms, KEGG pathways [76] and clusters of orthologous groups (COG) with their associated functional categories [81].

4.1.2 Homology analysis

A BLAST analysis was performed using aforementioned FASTA files as queries against a local database of 386 fungi, containing non-redundant sequences considering only one representative organism of each fungal species, with the exception of some *T. delbrueckii* closely-related species that had more than one strain. In particular, from the group of 386 organisms, 19 strains of *Torulaspora*, Zygotorulaspora and Zygosaccharomyces genera, were also annotated using YGAP, in order to allow their inclusion to search for homologies with *T. delbrueckii*. Thirty-five *S. cerevisiae* strains whose origins were related with winemaking or fermentative beverages were also included. *T. delbrueckii* COFT1 genome was selected as query since it was the only one with non-nuclear information, having a complete genome assembly,

was sequenced using both short and long-read technologies, and corresponds to a strain originating from winemaking environments [82], going in line with the objectives of the present work. We considered also, for comparison, the remaining three *T. delbrueckii* assemblies. In summary, in the BLAST analysis each query corresponded to the alignment of a protein coding sequence in *T. delbrueckii* COFT1 against the local database.

4.1.3 Phylogenetic analysis

The full proteome of *T. delbrueckii* COFT1 was used to search the common proteome portion of 386 fungi, in a total of 329 fungal species' defined proteomes, belonging to five phyla: Ascomycota, Basidiomycota, Chytridiomycota, Microsporidia and Mucoromycota. The 5001 proteins of *T. delbrueckii* COFT1 were used as BLAST queries in a local database containing the proteins of the other organisms. Following the BLAST searches, the proteins where representatives of the other 385 organisms were detected, were filtered.

Each set of probable homologous proteins (containing the query and the results obtained for that query) were multiple-aligned using Clustal Omega [83]. Following the alignments, all proteins from a given species were concatenated using the alignment results. With this approach was obtained the common proteome between the analyzed organisms (a core conserved proteome containing mostly essential genes not related with specific biological traits of each species) fully aligned.

The concatenated alignment was used for phylogenetic reconstruction using maximum-likelihood in IQ-TREE [84] with the JTT model of amino-acid evolution and gamma-distributed rates (four rates) with 200 bootstrap replicates. Figtree (http://tree.bio.ed.ac.uk/software/figtree) was used to visualize and edit the tree.

4.2 Newly sequenced genomes exploration (comparative analysis)

4.2.1 Strain cultivation and DNA extraction

Samples were collected in different points of the globe by members of the TODOMICS project. Yeasts were grown overnight in YPD medium (0.5% Yeast extract (w/v), 1% Peptone (w/v), 2% Dextrose (w/v)). Yeast DNA was extracted using the DNeasy PowerMax® soil kit (Qiagen), following the manufacturer's

instructions. DNA concentration was confirmed in all samples by Nano drop ND-1000 spectrometer. The collected samples were sent to Novogene to be sequenced by Next Generation Sequencing (NGS).

4.2.2 Data received

The data received were raw reads in FASTQ format. Those raw reads had already been filtered. The filtering process consisted on the removal of:

- ✓ Reads containing adapters.
- ✓ Reads with N > 10% (N corresponds to position where it is not possible to determine the base).
- ✓ Reads containing low quality (Qscore<= 5) base which is over 50% of the total base.

The 5' adapter sequence was 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3', and the 3' adapter sequence was 5'-GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG-3'

Figure 10 displays the detailed statistics of the sequencing data quality, that arrived with the data.

Sample	Library	Flowcell/Lane	Raw reads	Raw data(G)	Effective(%)	Error(%)	Q20(%)	Q30(%)	GC(%)
T1	FDMS192244642-1a	HTTY3DSXX_L2	5464208	1.6	99.95	0.03	97.50	92.62	39.36
T2	FDMS192244643-1a	HTTY3DSXX_L2	4996205	1.5	99.96	0.03	97.41	92.52	37.53
T3	FDMS192244644-1a	HTTY3DSXX_L2	6067466	1.8	99.94	0.03	97.32	92.15	40.06
T4	FDMS192244645-1a	HTTY3DSXX_L2	6689197	2.0	99.95	0.03	97.32	92.16	40.22
T7	FDMS192244646-1a	HTTY3DSXX_L2	6291719	1.9	99.94	0.03	97.60	92.84	40.29
T8	FDMS192244647-1a	HTTY3DSXX_L2	6053572	1.8	99.94	0.03	97.40	92.35	39.84
T9	FDMS192244648-1a	HTTY3DSXX_L2	7229168	2.2	99.95	0.03	97.31	92.16	40.28
T11	FDMS192244649-1a	HTTY3DSXX_L2	6511913	2.0	99.93	0.03	97.67	92.98	40.09
T12	FDMS192244650-1a	HTTY3DSXX_L2	5847316	1.8	99.92	0.03	97.33	91.93	29.02
T13	FDMS192244651-1a	HTTY3DSXX_L2	6101676	1.8	99.93	0.03	97.29	92.14	40.38
T14	FDMS192244652-1a	HTTY3DSXX_L2	6134636	1.8	99.93	0.03	97.70	93.06	40.04
T15	FDMS192244653-1a	HTTY3DSXX_L2	5817270	1.7	99.92	0.03	97.60	92.85	40.75
T19	FDMS192244654-1a	HTTY3DSXX_L2	6560673	2.0	99.93	0.03	97.37	92.37	38.98
T20	FDMS192244655-1a	HTTY3DSXX_L2	6322999	1.9	99.92	0.03	97.54	92.69	41.20
T22	FDMS192244656-1a	HTTY3DSXX_L2	5529555	1.7	99.94	0.03	97.76	93.20	41.00
T23	FDMS192244657-1a	HTTY3DSXX_L2	6271786	1.9	99.94	0.03	97.54	92.70	39.14
T26_2	FDMS192244658-1a	HTTY3DSXX_L2	6345665	1.9	99.95	0.03	97.57	92.72	39.64
T17	FDMS192244659-1a	HTTY3DSXX_L2	6555907	2.0	99.82	0.03	97.43	92.47	38.04

Figure 10. Data quality summary received with the sequenced data by Novogene.

4.2.3 Raw reads treatment

First of all, fastQC [53] was run for every strain data, like "fastqc *fq.gz". To check that the adapters had indeed been removed, grep was use to search for the sequence in the received files.

Figure 11. Pipeline followed to perform the treatment of the raw reads.

At this point, it was decided to perform an alignment with *T. delbrueckii* CBS 1146, the reference genome and, with *T. delbrueckii* COFT1, the only genome available with non-nuclear information. The chosen program to perform the alignment was BWA [64]. To perform the alignment first, it was necessary to index the reference genome. Once the reference genome was indexed, it can be called by BWA. In our case, the two reads files were called, with the parameters "mem" and "-M", in order to obtain the alignment in an output SAM file it was written ".sam" after the ">". The SAM file must be sorted; this was done using Samtools sort option [85]. Then the files can be indexed, also using Samtools, now the optin index [86]. Samtools can also be used to transform SAM files in BAM files. BAM files are in binary format, being therefore much smaller, and, sometimes this format is necessary to some analysis. To convert the SAM files to BAM, Samtools view [87] was used, with the option –S. Once again the indexing was performed, but now on the BAM files.

4.2.4 IGV

IGV 2.8.0 [66], [88] allows to introduce a reference genome and the assembled reads, permitting a visual inspection. Since our reference genome was not a "hosted genome", it was necessary to load it. To do so, the FASTA file can be loaded and, if it is the case, the annotation file can be introduced next. Another option, and the one followed in this work, is to create a genome file with the FASTA file and the other files of interest. Either way, it is mandatory to do the index of the FASTA file before [89].

Regarding the files of the strains we aligned, they had to be in the BAM format and the corresponding indexes files must be in the same directory. Once the information is loaded in IGV, data inspection can be done, from a visual evaluation of the alignment, to the study of individual genes.

4.2.5 VCF

Variant calling files are very important since they contain the information regarding the differences between a reference genome and one being studied and it is the fundamental method of choice to store human genomic data namely in clinical studies. VCF files can be generated with freebayes 1.3.3 [90], [91], to the parameter –f is given the reference genome and then is given the name of our BAM files. The output is a VCF in gz mode. This was done for each BAM file, obtaining, therefore, one VCF file for each of the newly sequenced strains. This process was performed for all the strains twice, one with *T. delbrueckii* CBS 1146, and another with *T. delbrueckii* COFT1 as reference genome.

4.2.6 Statistics

To better understand the data, statistical analysis can be done. Flagstat [92] from Samtools provide information such as properly paired reads, from the analysis of the BAM files. Flagstat was applied as can be seen in the next command line.

It is also possible to have a statistical analysis of the VCFs, with vcf-stats [93], from VCFtools, wich consists in a set of tools written in Perl and C++ designed for VCF files analysis. The VCF's previously produced with freebayes, were subject to vcf-stats and their information was summarized, and written in a .txt.

4.2.7 Consensus

The consensus sequences were generated with vcf-consensus using the VCF files previously generated and the correspondent reference genome. First of all "tabix –p vcf" was used and then "cat reference_genome.fna | vcf-consensus "${filename}"". The resulting output was a file with the consensus sequence.

To check if the files differed from the reference sequence, the next command line was run, it printed 'files are identical' or 'files are different'.

- cmp –silent reference_genome.fna Tn_consensus.fa && echo 'files are identical' || echo 'files are different'

Moreover, it was checked manually the first alteration for one of the cases. It was verified that the first alteration registered in the respective VCF file, was indeed in the consensus file obtained, and the reference sequence had the original version.

4.2.8 *De novo* Assembly

It was decided to carry out a *de novo* assembly, although a reference assembly was already performed, in order to have two types of genome assembly, allowing to compare and double check the data and the obtained information. Therefore, every strain was subjected to a *de novo* assembly with SPAdes genome assembler v3.11.1 [61], [94]. The obtained alignment was submitted at YGAP to predict coding protein genes.

4.2.9 PCA with all the available *T. delbrueckii* genomes

In the first phase a study was performed regarding strains available at NCBI, in a second phase strains collected by our group were sequenced and studied. Now, in a third phase, both type of data is used to perform a PCA analysis. Published samples were downloaded from NCBI and then aligned to the reference genome, *T. delbrueckii* COFT1, using miniclip2 [95]. The function "Full genome/assembly alignment" (-ax asm10) was used with the divergence level at 1% for better alignments. After all samples are align to the reference, duplicate reads were excluded using SAMtools [96], [97] markdup function and later samples had their unmapped reads removed, also using SAMtools. With the samples now filtered, they were converted into genotype array using pileupCaller (SequenceTools – pileupCaller [98]), with the final format being EIGENSTRAT. After this the array was converted to PLINK format where samples where further trims to remove unnecessary information. With the files PCA [99] analysis was done.

5 RESULTS AND DISCUSSION

5.1. Homology analysis and genome annotation

5.1.1 *Torulaspora delbrueckii* genome annotation

Genome annotation of *T. delbrueckii* using YGAP yielded between 4228 and 5016 putative coding sequences (CDS), as described in Table 1. In particular, the lowest number of CDS was obtained when considering strain NRRL Y-50541, even though it was the longest genome of the four considered (11.53Mb, in comparison with an average of 9.42Mb of the remaining three), and the highest value was obtained screening the genome of strain SRCM101298 (5016).

Table 2. *Torulaspora delbrueckii* genomes used in this study, and corresponding number of protein coding sequences (CDS) and transposable elements predicted by YGAP.

Strain	Source	Reference	Coding Sequences	Transposable Elements (TY)	Homologies with *S. cerevisiae*	Unidentified coding sequences[1]
CBS 1146	Unknown; type strain	[100]	4978	5	4514	464
COFT1	Wine fermentations	[82]	5001	5	4506	503
NRRL Y-50541	Mezcal-fermentations	[101]	4228	6	3875	486
SRCM101298	Fermented food		5016	7	4513	468

[1] No homologies with *S. cerevisiae* S288c detected by YGAP

A high homology was found by YGAP between *T. delbrueckii* genome and the one of *S. cerevisiae*, with an average value of 4352 protein coding sequences detected as homologous, corresponding to 90.52% of the total annotated genes. The highest number of homologies with the strain *S. cerevisiae* S288c was obtained with the genome of strain CBS 1146 (4514), the *T. delbrueckii* type strain, even though this number was very similar with the one obtained with COFT1 – 4506 – and with SRCM101298 – 4513.

To go further, BLAST was used to identify proteins which revealed no homology with *S. cerevisiae* S288c, and, in this way, were labeled as unidentified. In detail, the unidentified proteins (between 464 and 503 - last column of table 1) were used as query against the NCBI RefSeq database, and BLAST results (top

5 hits for each protein) were clustered considering the taxonomic groups with top-results. Figure 12 summarizes the results obtained, considering only species with more than 10 hits.

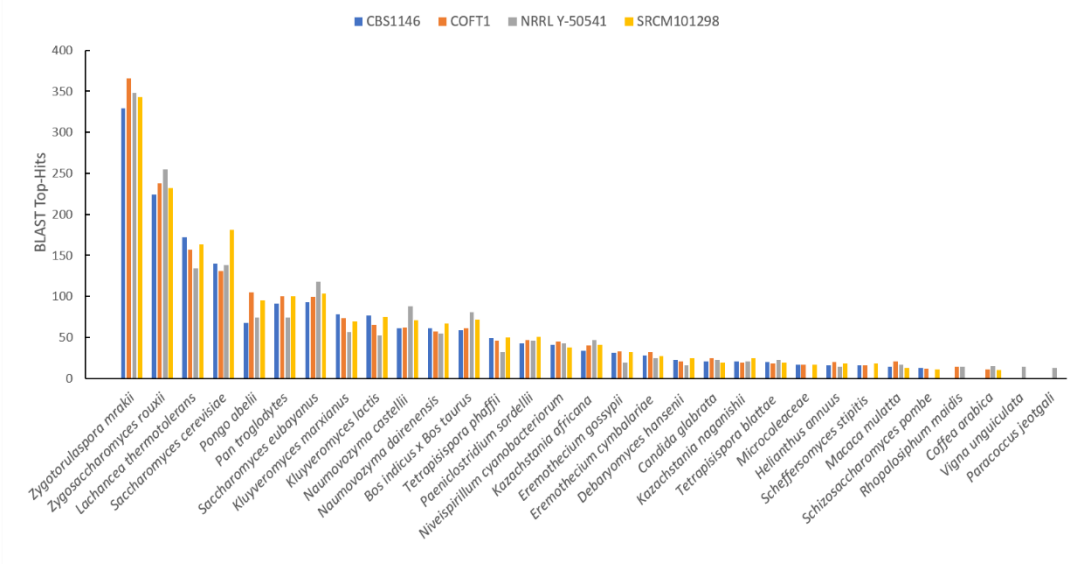


Figure 12. BLAST top-hits distributed by species, on the basis of best sequence alignments and lowest E-values, considering proteins not identified by YGAP in the four *T. delbrueckii* strains and five top-hits for each protein. Only species with more than 10 top-hits are shown.

Results showed that the higher percentage of unidentified proteins had a match with species *Zygotorulaspora mrakii*, as expected, followed by the genera *Zygosaccharomyces* and *Lachancea*. One surprising result was the homology detected between *T. delbrueckii* strains and two genomes of primates available in RefSeq database - *Pongo abelii* and *Pan troglodytes*. In fact, unidentified proteins matched with sequences from these genomes in higher proportion than the one found when considering genomes of other yeasts, such as the ones belonging to *Kluyveromyces*, *Naumovozyma* and *Candida* genera. More likely, these are not realistic matches but instead they are random matches when homologous genes were not detected in fungi (or at least not detected outside the main matches in *Zygotorulaspora* and *Zygosaccharomyces*), which was supported by no matches detected in BLAST against a fungi database in the following analyses. Our approach allowed for the first time to address and characterize *T. delbrueckii* full proteome, providing an important foundation for further studies exploring biotechnological uses of this species.

5.1.2 Functional annotation

EggNOG-mapper was used to perform functional annotation for the deduced *T. delbrueckii* proteins, offering insights into its biological significance (Figure 13). A total of 4814 genes of the *T. delbrueckii*

COFT1 genome (96.1% of the total annotated genes) were clustered by eggNOG-mapper in 24 clusters of orthologous groups (COG; Figure 13A), gathered in three main functional categories, as shown in panel 13B. Results show that the higher percentage of annotated genes are related with "information storage and processing" (28%) and "cellular processes and signalling" (27%). A high number of annotated genes didn't have a clear function attributed by egg-NOG (22%), and 23% of the genes were related with metabolism. The most abundant COG category with function attributed in the genome of *T. delbrueckii* COFT1 was "Intracellular trafficking, secretion, and vesicular transport" (485 genes, corresponding to 10.1% of the annotated genes), followed by "Transcription" (404 / 8.4%). The least abundant categories were "Cell motility" with only 1 gene clustered (0.02%), and "Nuclear Structure" (3 / 0.06%).

Small intra-species variability was found when considering functional annotation of the four *T. delbrueckii* available genomes (Figure 13C). Some exceptions were observed, in particular regarding genome of strain NRRL Y-50541, and COG categories "L: Replication, recombination and repair" and "I: Lipid Transport and Metabolism", for which a decrease in the number of genes in those clusters was detected (149 and 98, in comparison with 220 and 130 genes annotated in COFT1, respectively). COG data-base [102] has been a popular database for functional annotation of microbial genomes allowing the reliable assignment of orthologues to most genes [81]. Orthologous genes are products of speciation, and by being clearly defined, it allows to define relationships between species and to understand their evolution. The identification of orthologous genes was used previously to successfully identify differences and similarities between species, annotating their functional genetic information, proposing also functions in newly sequenced genomes [103].

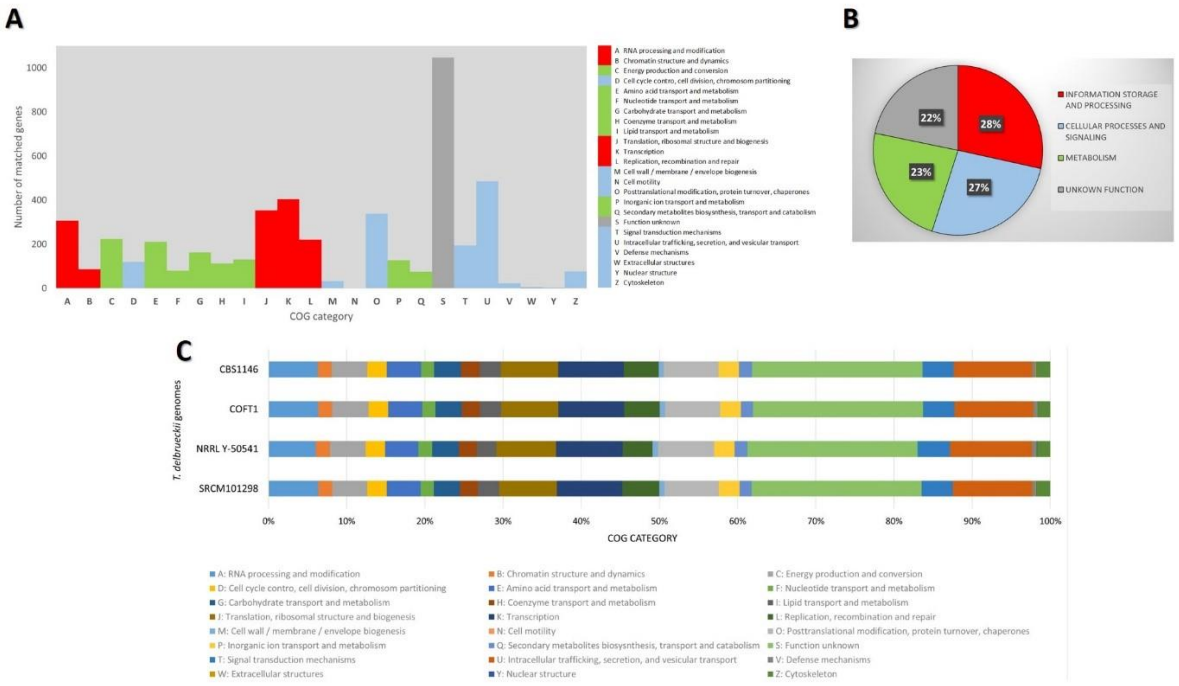


Figure 13 EggNOG classifications of annotated T. *delbrueckii* genes. Functional annotations were divided into 24 categories, correspond-ing to clusters of orthologous groups (COG). **A**: number of genes clustered in each of the 24 COG categories for the T. *delbrueckii* COFT1 genome. Colors are indicative of the functional categories used in panel B. **B**: Classification of T. *delbrueckii* COFT1 genes into functional categories. **C**: Comparison between the four available genomes of T. *delbrueckii* in terms of number of clustered genes (in percentage) in each COG category.

Kyoto Encyclopedia of Genes and Genomes (KEGG) [76], [104] was used to scrutiny eggNOG results interpreting the biological function of genes via interpretation of enzymes and biochemical processes. In the present study eggNOG-mapper allowed also to organize the 4814 genes in 3123 KEGG Orthology annotations. Table 2 represents the top results obtained in which at least four genes were grouped together under the same Orthology annotation. A considerable number of genes were assigned to "Protein-serine/threonine kinase" and "Amino acid transporters" (19 and 16 genes, respectively), being these the dominant categories. The high number of coding genes related with the phosphorylation of serine and threonine are in accordance with a high capacity of *T. delbrueckii* to consume these nitrogen sources detected already in fermentative trials, in particular in beer fermentation [105]. The capacity to effectively assimilate these aminoacids contributes to yeast growth and maintenance, but could function also as an important precursor for flavor formation and to improve fermentation fitness.

In *S. cerevisiae*, statistical differences were already observed in some strains regarding the consumption of several nitrogen sources, including serine and threonine, associated with different expression of genes involved in TORC1 pathway of nitrogen consumption [106]. Of notice is the fact that several genes were assigned to KEGG Orthology groups related with "Transporters". Knowledge about relevant aspects of

biology and biochemistry is still limited considering *T. delbrueckii*, including details about transport mechanisms and transporters collection, for example to assure the uptake of sugars during fermentation. In *S. cerevisiae* these transporters have a key role in the metabolism of carbon compounds [107], [108]. Regarding *T. delbrueckii*, recent results show a similar importance attributed to transporters [109], [110], stating the importance of this species to be used in food industry.

Table 3. Top KEGG Orthology and Pathways associated with *Torulaspora delbrueckii* COFT1 predicted protein coding sequences. Only categories with at least four genes were considered.

KEGG Orthology	KEGG Pathway description	Number of predicted coding genes
K08286	Protein-serine/threonine kinase	19
K16261	Amino acid transporters	16
K01509	Purine metabolism	8
K06867	uncharacterized protein	8
K08139	Meiosis	7
K00128	Glycolysis / Gluconeogenesis	5
K00129	Glycolysis / Gluconeogenesis	5
K00728	Mannose type O-glycan biosynthesis	5
K01210	Starch and sucrose metabolism	5
K01802	Metabolism	5
K07975	GTP-binding proteins	5
K08197	Transporters	5
K21989	Transporters	5
K00326	Amino sugar and nucleotide sugar metabolism	4
K00948	Pentose phosphate pathway	4
K01120	Purine metabolism	4
K01426	Arginine and proline metabolism	4
K01537	Metabolism	4
K03457	Transporters	4
K03854	Glycosyltransferases	4
K06883	function unknown	4
K07117	function unknown	4
K10967	Glycan biosynthesis and metabolism	4
K11121	Nicotinate and nicotinamide metabolism	4
K15109	Thermogenesis	4
K17741	Pyruvate metabolism	4
K19791	Transporters	4

5.1.3 Homology analysis

Upon obtaining and parsing the output from YGAP, as described in methods section, a BLAST analysis was performed to search for homologies between the selected coding regions of *T. delbrueckii* and the

NCBI genome database (database assessed in August 2020), in order for putative matches to be considered as homologous. Comparisons were analysed considering 386 yeast species with full genome sequences available in NCBI. From the 386 genomes, 329 corresponded to different species, and the remaining 57 to different strains of some species previously known to be closely related with *T. delbrueckii*, in order to obtain a greater detail of analysis. Results show the number of homologous protein coding sequences obtained for the totality of the organisms. Figure 14 summarizes the main results, representing BLAST top hits obtained for the species and genera with highest homology with *T. delbrueckii* detected: *Zygotorulaspora*, *Zygosaccharomyces*, *Lachancea* and *Saccharomyces*.

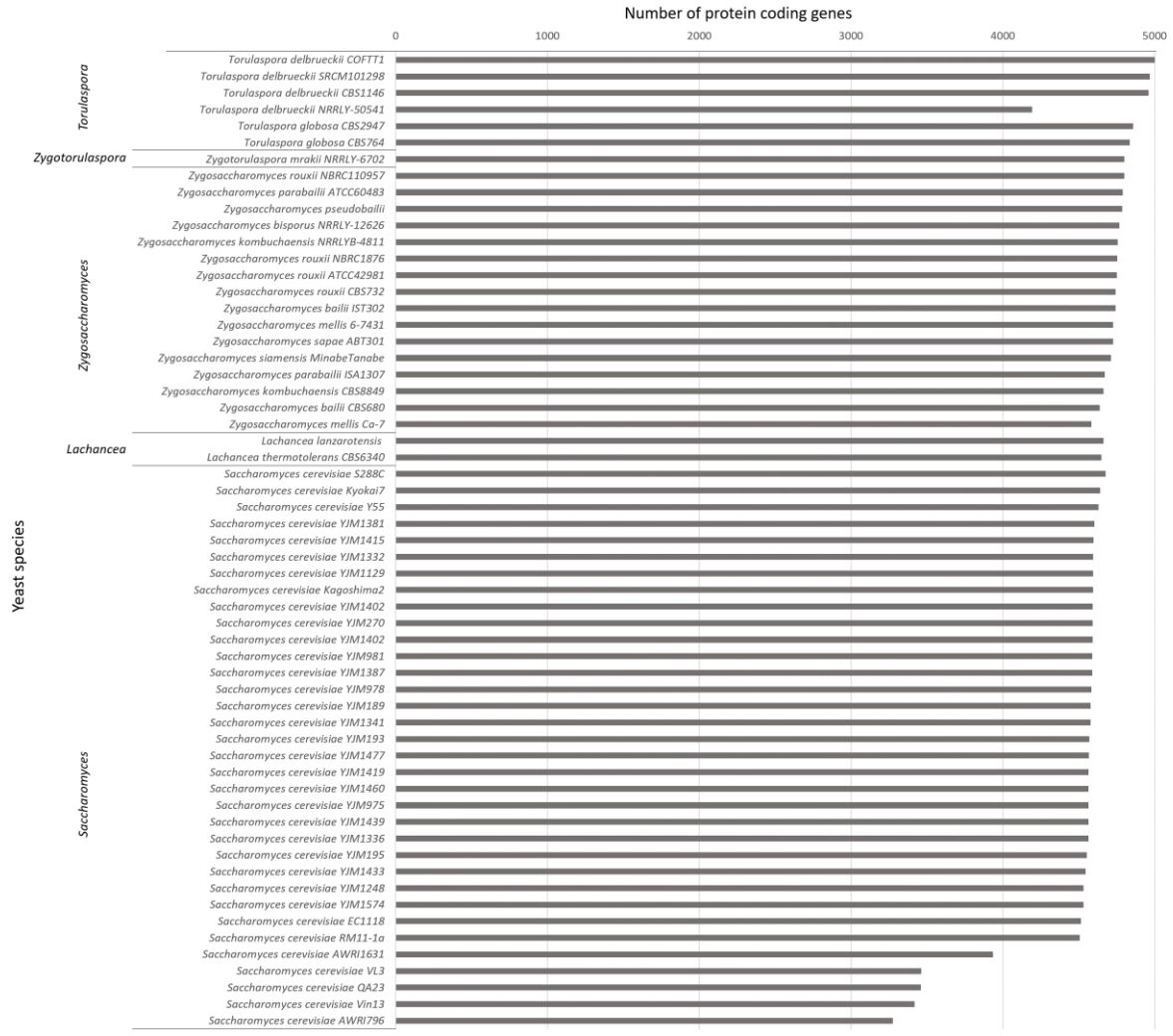


Figure 14. Homology comparison between protein coding genes of *Torulaspora delbrueckii* COFT1 genome (used as reference) and 56 related yeast species/strains. Protein coding regions of COFT1 genome were detected by YGAP and homology was determined by BLAST analysis.

Strain *T. delbrueckii* SRCM101298, originating from fermented food, obtained the higher percentage of homology with COFT1 (4969 common protein coding sequences out of 5001 used as query, corresponding to 99.4%), inside the group of *T. delbrueckii* strains, although very close to the one obtained for the type strain CBS 1146 (99.2%, 4960 common coding sequences). The three *T. delbrueckii* genomes shared 4960 homologous sequences, out of a total of 5001 putative coding sequences. The fourth genome considered – *T. delbrueckii* NRRL Y-50541 – revealed smaller homology, a fact in line with differences already found when analyzing FASTA files obtained from YGAP. We believe that these differences must not be due to true differences or lack of genome quality but, instead, to sequencing errors that lead to an absence of some parts of the genome.

When comparing *T. delbrueckii* with other species, the highest homology was detected in the genomes of *T. globosa*, as expected since they share the same genus. Strain *T. globosa* CBS2947 shared 4858 coding sequences with the genome of *T. delbrueckii* COFT1, corresponding to 97.1% of homology, while with strain *T. globosa* CBS764, 96.7% of homology was detected. The genera Zygotorulaspora and Zygosaccharomyces revealed 96% (4799 sequences) and 94.5% (4725 sequences, in average) of homology with *T. delbrueckii*, respectively. In particular, for Zygosaccharomyces species, between 4582 and 4800 putative coding sequences were revealed as homologous with *T. delbrueckii* COFT1, being *Zygosaccharomyces rouxii* NBRC110957 the most homologous strain (96.0%). Following, two species of Lachancea genus – *L. thermotolerans* and *L. lanzarotensis* – showed relevant amount of homology with *T. delbrueckii* COFT1 – 93.0 and 93.2%, respectively, which is in accordance to their role in fermentation, especially in fruit wine fermentation [111]. Surprisingly, the genus Saccharomyces, in particular the species *S. cerevisiae*, revealed slightly less homology with *Torulaspora* than the *Zygosaccharomyces* species. In fact, almost all *S. cerevisiae* strains (strain S288c was the only exception) revealed a smaller number of homologous genes (between 4506 and 4642 coding sequences) with *T. delbrueckii* genome than those obtained by the majority of Zygosaccharomyces species.

Within the group of Saccharomyces species, *S. cerevisiae* reference strain S288c showed the highest homology with the genome of *T. delbrueckii* COFT1, with 4676 homologous sequences (corresponding to 93.5%). Regarding other strains related with winemaking, from which, hypothetically, a higher homology would be expected, since *T. delbrueckii* COFT1 was originated in winemaking environments [82], this value was even smaller than the one obtained for the laboratory strain S288c. The lowest number of homologous putative coding sequences was obtained for the Australian strain AWRI796 (3276 – 65.5%), used worldwide as a commercial strain for winemaking (Mauri Yeast, Australia). Previous studies revealed

that this industrial strain, although showing a good fermentation performance, has particular genomic profiles mainly related with its extremely high sensitivity to harsh and stressful enological conditions [112], which could explain its distance from other winemaking strains considered in the present study. In fact, this connection between genetic features and their relevance in phenotypic variability and applicability in winemaking, was also shown before for other 172 *S. cerevisiae* wine strains [113], [114]. *S. eubayanus* and *S. paradoxus* revealed a similar level of homology to the average one obtained with *S. cerevisiae* - 91.5% and 92.8% -, respectively.

Genera *Naumovozyma, Kluyveromyces, Kazachstania* and T*etrapisispora* followed, in terms of decreasing order of homology with *T. delbrueckii*, with a smaller number of homologous sequences obtained (4580, 4517, 4531, 4469, respectively and in average). One case of particular notice was related with the genome of *C. glabrata*, that showed a total of 4492 hits (89.8%), a value similar to those obtained by *S. cerevisiae*, and different from the those obtained by other Candida species, which points to a possible proximity between these two species. This fact was also observed when the full yeasts' proteome was analyzed, as will be shown and discussed below.

Several studies have compared the fermentative potential of *S. cerevisiae* with the one of *T. delbrueckii*, although a full genomic comparison was still lacking, mainly due to the fact that *T. delbrueckii'* s available genomes were still sparsely annotated, and so, could not easily be compared with the well-annotated genome of *S. cerevisiae*. Although having marked differences at producing secondary metabolites during fermentation, as well as resisting to stresses, large similarities have been found at taxonomic and genetic levels between the two species. In fact, *T. delbrueckii* was previously identified as *S. rosei*, suggesting in this way a similar lineage to that of *S. cerevisiae*. In the present study, a total of 93.5% of homology was detected between genomes of *T. delbrueckii* COFT1 and *S. cerevisiae* S288c (Figure 14). Even though a higher score was expected due to the recognized proximity between Torulaspora and Saccharomyces genera, it is not surprising since Saccharomyces genus has evolved after a genome duplication event, in opposition to the *Torulaspora* and *Zygosaccharomyces* genera, that represent lineages that were separated from the *S. cerevisiae* one prior to whole genome duplication [115], [116]. The comparison between *S. cerevisiae* and *T. delbrueckii* was long discussed before, and regarding wine fermentation, *T. delbrueckii*, mainly due to their capacity to produce a different array of secondary metabolites, shows in fact a great potential to serve as an alternative to *S. cerevisiae*. A previous comparative analysis of transcriptome and metabolome of both species [15] detailed some important differences, in particular

the lack of multiple genes in *T. delbrueckii*, highlighting differences in the glycolic and fermentation pathways, together with a conclusion about a less volatile acidity associated with *T. delbrueckii*.

Homology was somewhat higher when the genome of *T. delbrueckii* was compared with the available genomes of *Zygosaccharomyces* species (between 94.5 and 96.0% - Figure 14). This percentage seems to indicate a proximity between *T. delbrueckii* and *Zygosaccharomyces* species, higher than the one found when comparing with *S. cerevisiae*, in terms of genomic analysis. However, by not being markedly different, the similarity points to a close proximity between the three genera, a fact already extensively discussed before, especially regarding physiological properties, but also using some genetic segments [15], [117]–[119]. In detail, Kurtzman and Robnett [120] using multigene sequence analysis compared 75 species belonging to the "Saccharomyces complex", including species of *Saccharomyces*, *Torulaspora* and *Zygosaccharomyces*. Species were divided into 14 clades, being species of genera *Torulaspora* and *Zygosaccharomyces* placed into three mixed clusters (7,8 and 9), apart from *Saccharomyces* species (both sensu stricto and sensu lato). To clarify these mixed clusters, authors have proposed in 2003 the creation of a new genus – *Zygotorulaspora* - comprising the species *Zygotorulaspora florentinus* and *Zygotorulaspora mrakii* [121]. Our results are in line with the conclusions of this work, since *Zygotorulaspora mrakii* NRRL Y-6702 showed 96.0% of homology with *T. delbrueckii*, a value higher than the one obtained for the genera *Zygosaccharo-myces* and *Saccharomyces*.

Zygosaccharomyces genus has been extensively studied over the years, and mostly associated with food spoilage. Especially the species *Z. bailii* and *Z. rouxii* have been often isolated as a contaminant during wine fermentation, mainly due to their high resistance to weak acids [122], [123]. However, and particularly in the last decade, the biotechnological potential of this genus has also been recognized [124], mainly for industrial bioprocesses involving low pH products or processes, high production of weak organic acids, heterologous proteins production, among others [125]–[128]. The similarity found when comparing *Zygosaccharomyces* genomes with those of *S. cerevisiae* strains can also be validated using previously obtained data. Mira et al. [129] described the genome of the acetic acid tolerant *Z. parabailii* ISA1307 strain, isolated from a sparkling wine production plant. Annotation of this genome revealed 4385 duplicated genes and 1155 predicted single-copy genes, mainly related with "metabolism and generation of energy", "protein folding", "modification and targeting" and "biogenesis of cellular components". It was concluded that genes related with these functions were also found in the genome of *S. cerevisiae* S288c, and in the one of *Z. rouxii* CBS732. Moreover, the most abundant motifs found in the proteins predicted in the analysis, revealed to be highly similar to the most abundant ones found in *S. cerevisiae*

S288c and *Z. rouxii* CBS732 genomes. This level of similarity is in accordance with the present work. Our results seem, in this way, to indicate that *Zygosaccharomyces* species form a group in between *T. delbrueckii* and *S. cerevisiae*, in terms of genomic comparison. This fact pinpoints the biotechnological potential of this genus, in line with recent discussions.

Of particular highlight is the homology detected for the genus *Lachancea* – 93.2% considering species *L. lanzarotensis* and 93% for species *L. thermotolerans*. This proximity between Lachancea and *T. delbrueckii* was shown before, detailing shared traits related mainly with osmotolerance and ethanol resistance [130], [131]. Especially *L. thermotolerans* have associated a strain-dependent production of a diverse range of metabolic intermediates for L-lactic acid production [131], [132], and also of ethyl lactate [133]. The high homology detected in the present study between the two species are also in line with their common capacity to ferment maltose, producing significant amounts of acetyl esters and long-chained ethyl esters [105], pointing in this way to the potential use of *T. delbrueckii* for industrial beer fermentation.

5.1.4 Phylogenetic analysis

A local database was compiled using 386 defined fungal proteomes, and compared against *T. delbrueckii* COFT1. Only yeasts having the full proteome characterized and annotated were considered (database built in August 2020). The entire proteome of *T. delbrueckii* COFT1 (5009 proteins) was used to BLAST against the database and a total of 204 *T. delbrueckii* proteins displayed homologues in the 386 fungi. A phylogenetic analysis (Figure 15) was performed considering the alignment of the core concatenated proteins present in the 386 organisms.

Our results display a general evolutionary relationship between strains independently of specific physiological adaptations of the species. To our knowledge, it is the first time that this analysis is performed considering such a high number of organisms and assessing their common proteome. Globally, results show that the core group of 204 common proteins allowed to separate between the five phyla of fungi – Ascomycota, Basidiomycota, Chytridiomycota, Microsporidia and Mucoromycota – revealing a close proximity between the phyla Ascomycota and Basidiomycota, that represent a sister clade to the one containing the phyla Chytridiomycota and Mucoromycota. These four phyla show a marked distance to the Microsporidia that represent a deeper split within the fungi group, only with the species *Mitosporidium daphnia* – showing a proximity to the four phyla. Considering fungal subdivisions, the core proteome reveals a clear distinction between the seven taxa of Ascomycota and Basidiomycota

(colored boxes). This group of core proteins appear to be conserved across all fungal species, even considering the ones more distant phylogenetically, as for example the four species of genus Encephalitozoon - *E. romaleae, E. hellem, E. intestinalis* and *E. cuniculi* -, belonging to class Microsporidia, and being the most common pathogenic genus of humans and domesticated animals in this class. Importantly, all phyla and major taxonomic groups within the phyla established monophyletic clades attesting for the robustness of the tree.

A more detailed analysis of *T. delbrueckii* placement (highlighted in detail in Figure 15), confirms the species as phylogenetically closer to Zygosaccharomyces species than to *S. cerevisiae*, as shown in our previous analysis. Our results are in accordance with the work of Shen et al. [134], showing the phylogenetic placement of more than 300 budding yeasts, and highlighting some genetic distance between *Torulaspora* and *Saccharomyces* genera, in favor of other more genetically closed genera such as *Zygosaccharomyces* and *Zygotorulaspora*. Our work represents an advancement of knowledge, including several other fungal species, in addition to the budding yeasts studied.

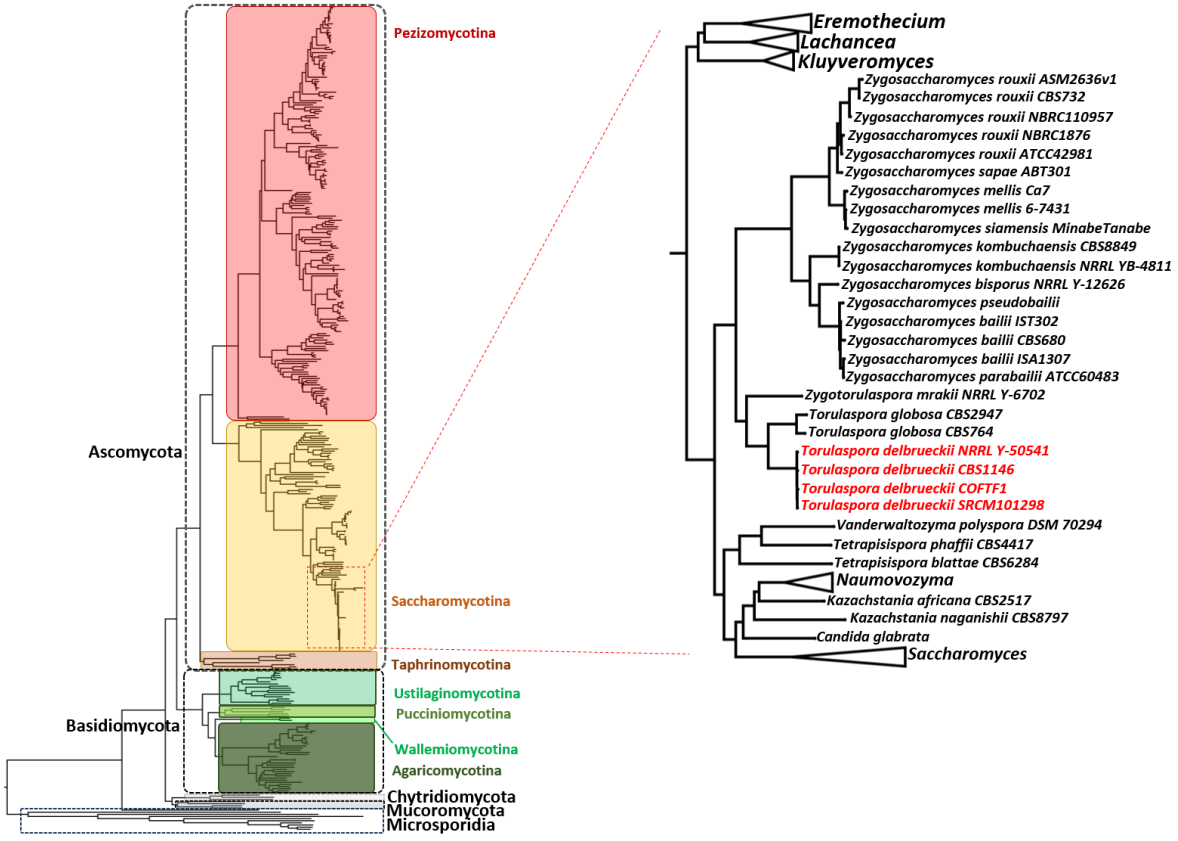


Figure 15. Phylogeny of fungi, considering 386 fungal core genomes (alignment of 204 common proteins). Phyla are highlighted using dashed lines, and subphyla are identified according to colored boxes. The placement of *Torulaspora*

delbrueckii strains are shown in detail in relation with the closely related species inside Saccharomycotina subphylum. The concatenated alignment was used for phylogenetic reconstruction using maximum-likelihood and 500 bootstrap replicates.

It is clear, when analyzing the phylogenetic tree of Figure 15 that the two *Z. rouxii* strains are closer between them, and then grouping with a clade containing two *Z. parabailii* and two *Z. bailii* proteomes, being the *Zygosaccharomyces* genus clearly the second most closely related with *T. delbrueckii*, just behind the *Zygotorulaspora* genus, considered as a sister genus. *S. cerevisiae* strains appear only further away in a group composed by other species from the genera *Kazachstania, Naumovozyma, Tetrapisispora* and *Vanderwaltozyma*. Importantly, and contrarily to what was concluded in the homology analysis, *Lachancea* spp. were located in a separated branch to the one containing *Saccharomyces* and *Torulaspora* species. Results show that, although *Saccharomyces* and *Torulaspora* species are evolutionarily closer, *Lachancea* and *Torulaspora* have a higher biochemical and physiological proximity, as shown by the higher number of homologous genes, as already discussed.

Also of notice is the fact that *C. glabrata* was located in the *Saccharomyces* group, close to *S. cerevisiae* wine strains and to *S. paradoxus* and *S. eubayanus*, and apart from the *Candida* subclade. This result is in accordance to what was shown before [135], discussing the similarity between *C. glabrata* and *S. cerevisiae*, although the first has evolved to acquire pathogenicity in mammalian hosts.

5.2 *T. delbrueckii* genome sequencing: comparative analysis and biotechnological potential assessment

In a second phase of this thesis, a study with 54 *T. delbrueckii* strains whose genomes were sequenced in NOVOGENE® facilities, was conducted. The aim of this analysis was to understand the relationship between *T. delbrueckii* strains, and to relate genomic information with the strains' origins and technological uses.

5.2.1 Visualization of yeasts genome using IGV

The BAM files, resulting from the alignment with BWA, were inputted in IGV, to have a visual analysis of the data. From the visualization of the alignments with *T. delbrueckii* CBS 1146 at IGV it was observed that the data could be grouped in 4 groups. Group 1 composed by T3, T5, T14, T44, T57, T58 and T59 that seemed very similar between them but with a high rate of alterations regarding the reference genome.

Group 2 constituted by strains T4, T7, T20 and T42, that unexpectedly revealed as very similar, both between them and with the reference genome. Group 3 composed by T37, T52, T53, T54, T61 and T62, that present profiles that allow to say those samples are not *T. delbrueckii* species. The fourth group consisted of the remaining strains that appeared similar between them, having some spaced differences.

In a general way strains from Group 1 had the higher lack of information in the extremities of the chromosomes and a higher rate of alteration along the chromosomes. The most outstanding case was the beginning of chromosome 5, that had lack of information until 46Kb. An example of that high rate of differences from the reference genome is in figure 16, that captures a section in the middle of chromosome 4, for two strains belonging to Group 1.

Figure 16. IGV capture of strains T3 and T14 chromosome 4, with *T. delbrueckii* CBS 1146 as reference genome.

Regarding Group 2, that was the group of strains whose assemblies had the lower rate of alterations observed in their extension, inclusively in the extremities. From that is possible to conclude that those strains are much more alike *T. delbrueckii* CBS 1146, the reference genome, than any of the other strains.

Figure 17. Capture of IGV visualization of the middle of chromossome 4 for T14, T8 and T7 strains, with *T. delbrueckii* CBS 1146 as reference genome.

When the analysis was made with *T. delbrueckii* COFT1 strain as the reference genome, the same groups were identified. Figure 18 is an illustrative image with three strains (T7, T8 and T14) as example.

Figure 18. Capture of the visualization of the middle of chromossome 4 for T14, T8 and T7 with IGV, being *T. delbrueckii* COFT1 the reference genome.

Comparing the same analysis with the two reference genomes, *T. delbrueckii* CBS 1146 and COFT1, the main difference observed in the data visualization was the apparent increase of alterations in T4, T7 and T20 strains, when the reference genome was *T. delbrueckii* COFT1. That means that those three strains are genetically very close to *T. delbrueckii* CBS 1146, corresponding, possibly to the same strain.

5.2.2 VCF statistical analysis

In order to have a statistical inference on the alignments from the bam files, it was necessary to obtain a statistical evaluation of the output for each strain. Flagstat, from samtools, is one tool that allows such information summarization. Samtools flagstat allowed to estimate, among other information, the QC-passed reads and how many reads were properly paired. The analysis was performed for all the sequenced strains, using *T. delbrueckii* CBS 1146 and *T. delbrueckii* COFT1 as reference genome. The first analysis had *T. delbrueckii* CBS 1146 as the reference genome, and after gathering the information, T12, T17, T32, T37, T52, T53, T54, T61 and T62 stood out, for having very low percentages of properly paired reads. Although having an average number of QC-passed reads, the percentage of properly paired reads, in those samples, were dramatically low, having values between 22.01% and 3.58%. For the remaining strains the percentage of properly paired reads was between 96.18% and 71.28%. Those were the same strains that also revealed a bad profile when analyzed with IGV. Therefore, it was concluded that those nine strains were not *T. delbrueckii* species.

Regarding the analysis with *T. delbrueckii* COFT1 as reference genome, the values of properly paired reads were higher, varying between 94.62% and 98.83%. This might be due to the fact that only *T. delbrueckii* COFT1 genome contains non-nuclear information. Since the samples from our collection were subjected to a whole-genome sequencing process, this means that among the reads we have is not only the nuclear information, but also the non-nuclear information. That being said, it is assumed that when *T. delbrueckii* COFT1 strain is used as a reference genome to perform the assembly, the percentage of properly paired reads is higher once with this reference we also have the non-nuclear information being aligned. In table 3 is summarized the information obtained from the flagstat analysis of all the sequenced strains.

Table 4. Summary of some outputted information by flagstat for the 16 strains, with *T. delbrueckii* CBS 1146 and COFT1 as reference genome.

Strains	*T. delbrueckii* CBS 1146		*T. delbrueckii* COFT1	
	QC-passed reads	properly paired	QC-passed reads	properly paired
T01	10952228	81.85%	10958341	97.75%
T02	10005814	71.28%	10024163	96.59%
T03	12154471	86.85%	12195505	94.77%
T04	13386579	87.93%	13397620	97.92%
T05	8899492	80.10%	8952864	93.40%
T07	12591021	88.43%	12600930	98.14%
T08	12120937	84.96%	12126926	98.27%
T09	14474296	87.18%	14478194	98.27%
T10	7496537	74.02%	7490765	86.78%
T11	13037824	84.61%	13040631	95.42%
T12	11699130	3.58%		
T13	12218750	87.28%	12224196	98.67%
T14	12288963	85.69%	12353923	94.62%
T15	11647012	90.05%	11646403	98.83%
T17	13114798	9.08%		
T19	13140814	80.93%	13148088	97.99%
T20	12657108	94.17%	12665771	97.94%
T22	11075927	91.59%	11078904	98.41%
T23	12558869	81.98%	12569778	97.52%
T26	12705093	85.21%	12716070	97.86%
T27	10919239	86.03%	10925463	98.84%
T28	10859743	88.97%	10864466	97.89%
T30	6357215	85.24%	6361839	98.11%
T32	11433411	17.79%		
T34	9459526	85.13%	9461114	98.23%
T35	9666975	88.37%	9669322	99.06%
T36	10331859	77.46%	10344887	97.24%

Strains	*T. delbrueckii* CBS 1146		*T. delbrueckii* COFT1	
	QC-passed reads	properly paired	QC-passed reads	properly paired
T37	11340135	6.83%		
T38	10412677	81.58%	10419683	98.95%
T39	10110650	84.78%	10113543	98.38%
T40	8957864	85.96%	8952525	99.30%
T41	12230566	83.77%	12229092	99.31%
T42	8549718	96.18%	8553909	97.65%
T43	8552056	87.02%	8559909	98.41%
T44	7145621	88.43%	7175309	95.47%
T45	8005989	90.82%	8006463	98.22%
T46	10310501	81.43%	10321607	94.21%
T47	7388081	85.69%	7387549	98.53%
T49	8139297	89.18%	8143369	95.40%
T50	7995340	85.36%	8000646	98.75%
T51	7805159	91.71%	7805908	98.41%
T52	8953603	9.47%		
T53	7288097	22.01%		
T54	8537404	20.84%		
T55	10354679	76.20%	10364915	82.19%
T56	9391182	87.55%	9389989	98.92%
T57	8077492	86.66%	8137884	96.30%
T58	6168930	89.71%	6186286	95.13%
T59	6814199	90.05%	6844415	96.29%
T60	10075524	77.35%	10080697	96.51%
T61	12440228	11.49%		
T62	8481862	10.42%		
T63	9559403	89.01%	9565052	98.04%
T64	8712942	90.13%	8720301	98.36%

VCF files were generated after strains genomes' alignment, using as reference the genomes of *T. delbrueckii* strains CBS 1146 and COFT1. In order to make a comparative analysis of the vcf's information, vcf-stats was used to analyze the VCF files of each genome and obtain a statistical analysis summary for each one. Among the statistics obtained with vcf-stats, the SNPs (single nucleotide polymorphisms) counts was the most relevant, together with the total number of indels (insertions and deletions). The analysis was performed for all the strains available, using *T. delbrueckii* CBS 1146 and COFT1 as reference.

The studied strains can be separated in three groups similarly to what was obtained in the IGV analysis. The group constituted by T4, T7, T20 and T42, present lower rate of SNP's and indels than the average, when the reference genome is *T. delbrueckii* CBS 1146. When *T. delbrueckii* COFT1 was used as the reference genome, the number of SNP's and indels was higher. Regarding T3, T5, T14, T44, T57, T58 and T59, the number of identified SNP's was around 300,000 and the number of indels around 8,000, for both reference genomes.

Table 5. Summary of some outputted information by vcf-stats for the 16 strains, with *T. delbrueckii* CBS 1146 and COFT1 as reference genome

Strain	CBS 1146		COFT1	
	snp_count	indel_count	snp_count	indel_count
T1	28339	2711	23602	2990
T2	24550	2458	27204	3119
T3	304315	9022	304833	9446
T4	1283	1464	30485	3252
T5	302276	8889	302755	9317
T7	1267	1447	29294	3182
T8	22534	2288	27940	3079
T9	22539	2338	27934	3109
T10	28601	2724	194	1714
T11	22045	2340	27171	3095
T12	7742	53		
T13	22812	2344	27583	3120

Strain	CBS 1146		COFT1	
	snp_count	indel_count	snp_count	indel_count
T14	302138	9033	302907	9501
T15	28163	2696	18867	2745
T17	134028	1082		
T19	27708	2686	18562	2703
T20	1303	1452	29395	3184
T22	22374	2363	27418	3121
T23	22244	2271	27886	3084
T26_2	28964	2653	26833	3102
T27	29572	2748	27420	3159
T28	29879	2741	27214	3107
T30	29995	2702	26548	3073
T3030	29996	2688	26531	3071
T32	31021	2735	27872	3083
T34	29862	2765	20068	2819
T35	28541	2732	16637	2562
T36	29322	2764	19959	2816
T37	27500	177		
T38	28392	2709	16462	2581
T39	29789	2793	18529	2747
T40	30044	2793	17867	2645
T41	29909	2803	18482	2737
T42	1354	1442	29147	3157
T43	29344	2662	26767	3143
T44	303743	8957	304371	9418
T45	24028	2346	28650	3105
T46	27362	2559	27078	3138
T47	28625	2719	20126	2842
T49	22641	2365	28905	3131
T50	30255	2726	27441	3114
T51	31741	2769	28293	3153

Strain	CBS 1146		COFT1	
	snp_count	indel_count	snp_count	indel_count
T52	191251	1406		
T53	198979	1464		
T54	202170	1494		
T55	44800	2502	51447	3301
T56	28669	2729	20217	2846
T57	233637	8564	233818	8953
T58	302843	8976	303120	9451
T59	234127	8546	234020	8905
T60	31490	2790	28287	3167
T61	81600	516		
T62	77500	451		
T63	29635	2666	26608	3088
T64	29707	2720	26513	3160

5.2.3 *De novo* assembly

In order to have a different perspective on the alignment, de-Novo Assembly was performed using Spades for all the newly sequenced *T. delbrueckii* strains. By the analysis of the metrics, summarized in table 3, it was possible to have a preview about the success of the assembly, mainly because the obtained predicted genome length was similar to the *T. delbrueckii* genomes available at NCBI, and used in the first part of this work.

Table 6. *De novo* assembly statistics, after analysis of the *T. delbrueckii* alignments with spades.

	Total length	Number of scaffolds	N50	#N's	Number of scaffolds > 1000pb
T01	9,261,029	232	1,056,173	1,130	48
T02	9,291,302	270	916,526	620	42
T03	9,220,330	111	1,067,218	800	34
T04	9,362,018	448	1,070,874	810	41
T05	9,185,765	109	1,067,912	600	36

	Total length	Number of scaffolds	N50	#N's	Number of scaffolds > 1000pb
T07	9,252,126	135	1,070,840	1010	34
T08	9,246,637	219	894,074	400	44
T09	9,271,045	231	1,052,208	800	43
T10	11,361,494	243	1,054,428	1,100	97
T11	9,296,892	244	847,160	800	56
T13	9,228,010	246	1,058,188	600	42
T14	9,295,047	142	1,053,496	1,000	42
T15	9,215,907	173	928,894	900	51
T19	9,216,474	135	1,031,427	810	43
T20	9,256,214	125	1,070,774	1,100	33
T22	9,222,144	187	1,053,002	900	38
T23	9,266,180	204	1,061,816	700	45
T26	9,251,466	153	879,459	900	52
T27	9,256,772	124	1,043,385	500	46
T28	9,286,028	158	849,787	800	57
T30	9,252,286	185	778,839	600	55
T34	9,255,778	133	837,514	810	51
T35	9,206,680	130	1,056,677	1,100	41
T36	9,249,569	184	953,266	1,120	53
T38	9,217,393	133	998,157	700	48
T39	9,252,866	166	1,029,332	900	48
T40	9,240,762	127	702,811	1,200	47
T41	9,217,557	134	999,509	810	46
T42	9,234,762	133	1,070,843	1,100	40
T43	9,236,457	199	999,764	710	58
T44	9,244,223	157	1,069,946	400	42
T45	9,256,361	206	882,304	610	47
T46	9,279,812	280	556,041	600	69
T47	9,250,800	204	631,845	1,100	51
T49	9,278,402	117	1,055,583	800	41
T50	9,252,126	146	1,030,614	900	45
T51	9,263,857	190	632,330	700	45
T56	9,248,272	201	984,357	1,000	52
T57	9,206,633	129	1,066,611	900	46
T58	9,273,894	139	1,072,188	900	46
T59	9,232,121	198	1,068,870	600	43
T60	9,371,680	308	705,274	900	56
T63	9,237,520	140	1,058,437	900	45
T64	9,249,089	181	845,129	800	55

Those alignments were later submitted to YGAP in order to obtain the same type of information as obtained in the first stage of this study, using genomes available in NCBI. The resulting information is summarized in table 4, and is in accordance to what was expected.

Table 7. Coding sequences predicted by YGAP for the alignments performed with SPADES.

Strain	Coding Sequences	Strain	Coding Sequences
CBS 1146	4978	T36	4931
COFT1	5001	T37	5384
NRRL Y-50541	4228	T38	4909
SRCM101298	5016	T39	4904
T01	4945	T40	4885
T02	4966	T41	4902
T03	4952	T42	4965
T04	4959	T43	4906
T05	4959	T44	4969
T07	4943	T45	4956
T08	4947	T47	4956
T09	4920	T49	4927
T10	5094	T50	4947
T11	4950	T51	4959
T13	4952	T52	5085
T14	4943	T53	4927
T15	4901	T54	4941
T19	4935	T55	9865
T20	4978	T56	4948
T22	4940	T57	4869
T23	4973	T58	4940
T26	4954	T59	4956
T27	4939	T60	4924
T28	4966	T61	5001
T30	4946	T62	4963
T34	4918	T63	4956

Strain	Coding Sequences	Strain	Coding Sequences
T35	4929	T64	4959

5.2.4 PCA with all the available *T. delbrueckii* genomes

In order to compare the genomes deposited at NCBI and studied at the first phase of this work and the new genomes obtained by us in the scope of this work, a PCA (Figure 19) was performed, taking into account all the available *T. delbrueckii* genomes. At this point ten new *T. delbrueckii* assemblies (L09; L10; L11; L12; L13; L15; L16; L18; L19; L20) [136] were deposited at NCBI and the opportunity was taken to perform this analysis with more strains and therefore have a more meaningful result.

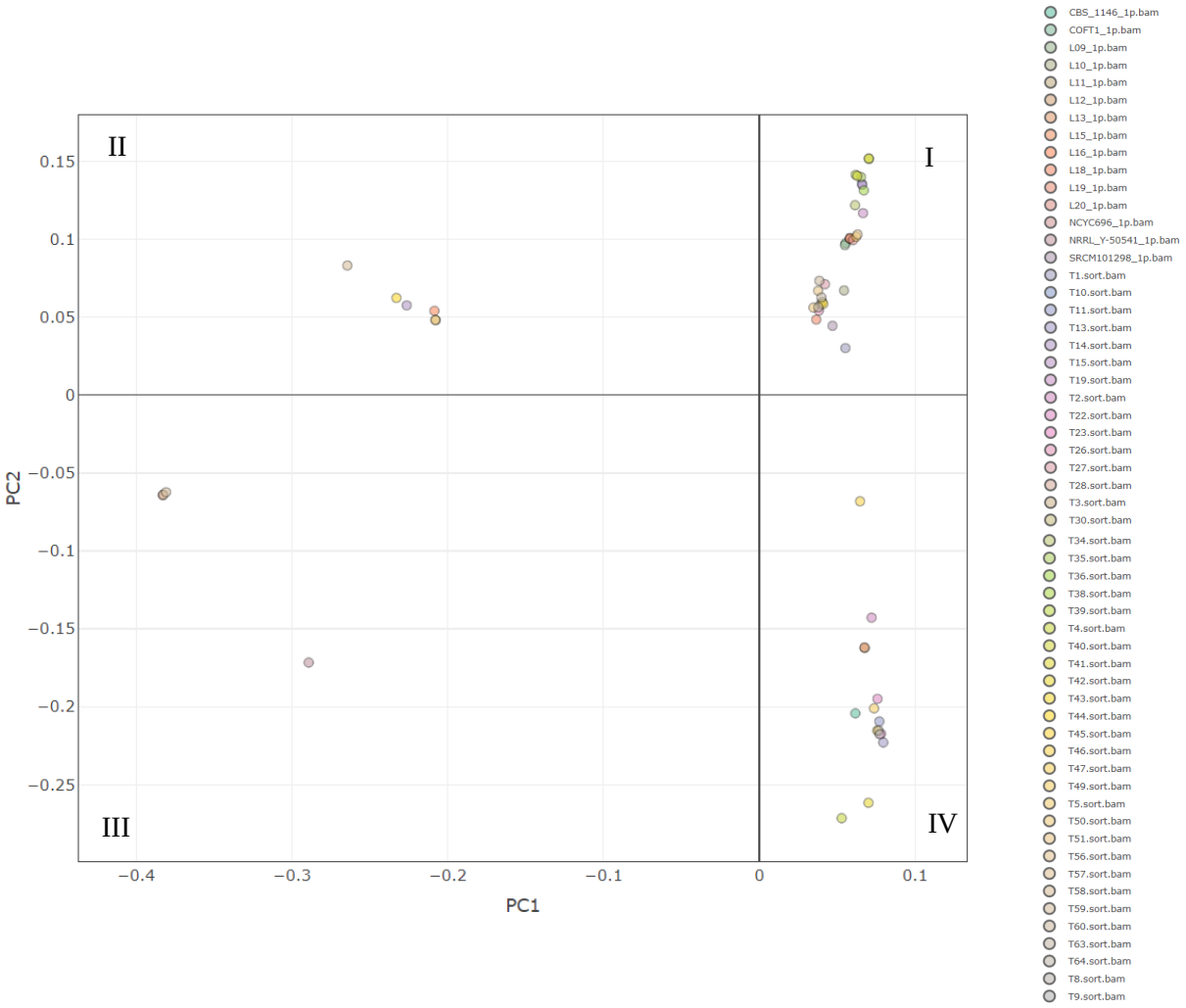


Figure 19. PCA with all the *T. delbrueckii* genomes available (from our collection and the public database).

Results of the PCA analyzed shows that all the considered genomes were grouped along the four quadrants of the PCA graphic, being observed the presence of some highly related clusters among them. Considering the data obtained in the PCA, we can conclude that the samples analyzed, both those extracted from the NCBI and isolated by our group, are divided into the groups as shown in the table below.

Table 8. *T. delbrueckii* strains grouped by the PCA.

Quadrants	*T. delbrueckii* strains
I	T1; SRCM101298; L18; T26; T64; T50; T28; T43; T30; T63; T51; T27; T60; L10; L09; COFT1; L20; L19; T47; T56; T19; T34; T10; T15 (T10 and T15 highly similar); T36; T35; T39; T40; T41; T38 (T41 and T38 highly similar)
II	T58; T44; T14; L16; T3; T5; (being T3 and T5 highly similar)
III	NRRL_Y-50541 NCYC696; T57; T59 (highly similar)
IV	T46; T2; L13; L12; T22; CBS_1146; T49; T45; T11; T8; T23; T9; T13; T42; T4 (being the following pairs highly similar; T8/T45, T9/ T23 and L12/ L13).

The strains observed at quadrant II do not appear to have a pattern that connects them. Regarding the strains at quadrant I, after crossing the obtained information in the PCA with the collection data, it is observed that the strains here grouped were mainly collected from winemaking. All the strains obtained from wine making from Portugal (isolated from grape must of Portuguese wine Castelão), and from winemaking from Spain (isolated from grape must of Prieto Picu and grape must of Tempranillo) are here clustered, emphasizing they are close, not only phenotypically but also at a genomic level. Also *T. delbrueckii* COFT1, gathered from NCBI, and registered as collected from must, is grouped close to those wine related strains. The only wine strains from our collection that are not present in this group are strains T22 and T23, that appeared in group IV.

Considering the strains at group III, T57 and T59 are both from SW Ontario, Canada. Both were isolated from Natural environments, Bark of Quercus rubra and Bark of Quercus velutina, respectively. NCYC696 was isolated from souring figs. The last three mentioned strains present a very high similarity between them. At last, NRRL Y-50541, that is further away from the remaining ones, was isolated from wort for Mezcal production.

Regarding the strains grouped at IV, those are mainly from food, bakery and even other beverages. T8 and T9 are the only two strains in our collection collected from Bakery, and are clustered together here. Close to them are also T22 and T23, the only two winemaking strains that are not in group I. In group IV are also gathered almost all the strains collected from a food subtract such as; potato sarch factory, fruit and vegetables (green beans and artichoke) and Cheese. In this category, only the strain collected from strawberry is not grouped with the remaining ones, being positioned at quadrant II.

In the "Natural environments" category all the collected strains grouped together at quadrant I, with the exception of three strains. These three strains are from Canada and Japan. The Japanese one can be found at quadrant II, and the two Canadian ones at quadrant III. As "Other beverages" we only had two strains in our collection. Rhagi from Indonesia is positioned at quadrant II and sorghum brandy (kaoliang-chui), from China atquadrant IV.

6 CONCLUSIONS AND FUTURE PERSPECTIVES

Torulaspora delbrueckii is a non-*Saccharomyces* yeast many times referred as an alternative to *Saccharomyces cerevisiae*, especially in wine and bread fermentations, contributing with a novel palette of aroma and flavor characteristics to the final product. The basis of this novelty has largely been searched, and genomic fingerprints of *T. delbrueckii*, exclusively found in this species, are believed to be interconnected with this question. However, the genome of *T. delbrueckii* being sparsely annotated, especially when comparing with the perfectly annotated genome of *S. cerevisiae*, doesn't allow to draw conclusions about the particularities of this fermentative yeast.

The present work represents a successful effort to increase and improve annotation of *T. delbrueckii's* genome, identifying homology between this yeasts and hundreds of other fungal species, together with a functional annotation of their coding genes, increasing their biological significance. Overall, this work provides a starting point to unravel the diversity of potential biotechnological applications of *T. delbrueckii*.

Regarding the genomes collected by our group, sequenced and studied here for the first time. Their genomes were aligned and *ab initio* annotation was achieved with YGAP. Following the mindset of Mudge et al. (2016) [74], which defines computational annotation as a process based on three methods: alignment, comparative annotation, and *ab initio* annotation, our intention in future work are to do the comparative annotation and improve *ab initio* annotation, for every strain. Moreover, in the future, we intend to continue this work by disclosing the ploidy level of the 54 strains with a reliable tool, achieve a phylogenetic analysis with the strains from our collection, perform sNMF and explore the genomic alterations that are in the basis of the characteristic phenotypes.

The information obtained in the present work will be of high relevance, as it has been described that ploidy and duplicated regions can have an impact on the fermentation performance of yeasts. Furthermore, annotation obtained with SPADES in the current work and submitted to YGAP, will follow the protocol of the first part of this work. Data will be of great importance to explore the fermentative potential of *T. delbrueckii* as wine yeast, serving as an alternative to *S. cerevisiae*.

7 ATTACHMENTS

7.1 IGV images

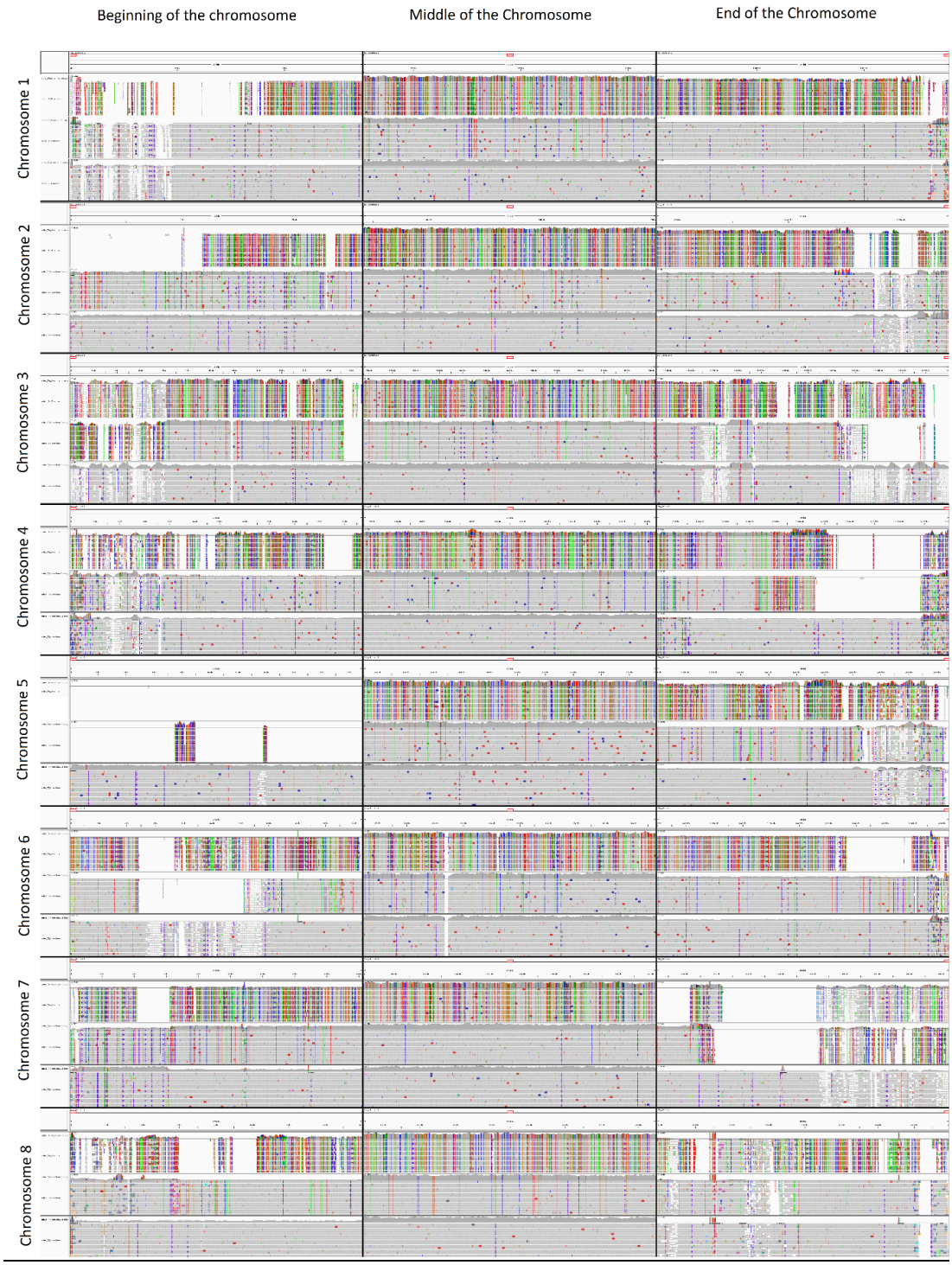


Figure 20. Representation of the visual visualization of T7, T8 and T14 *T. delbrueckii* strains when aligned with BWA and *T. delbrueckii* CBS 1146 as the reference genome.

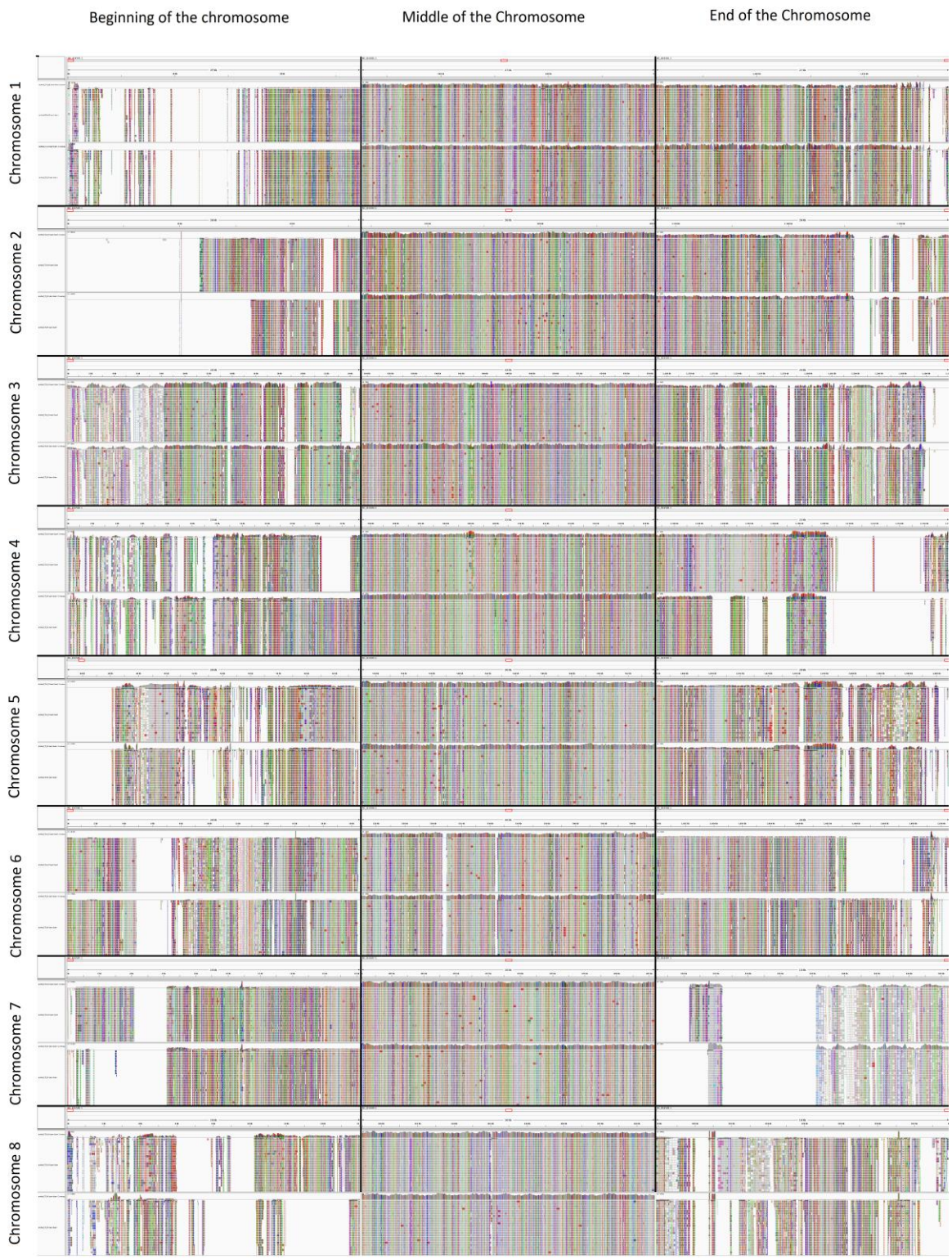


Figure 21. Visual representation of *T. delbrueckii* T14 and T3 strains when aligned with BWA and with *T. delbrueckii CBS* 1146 as the reference genome.

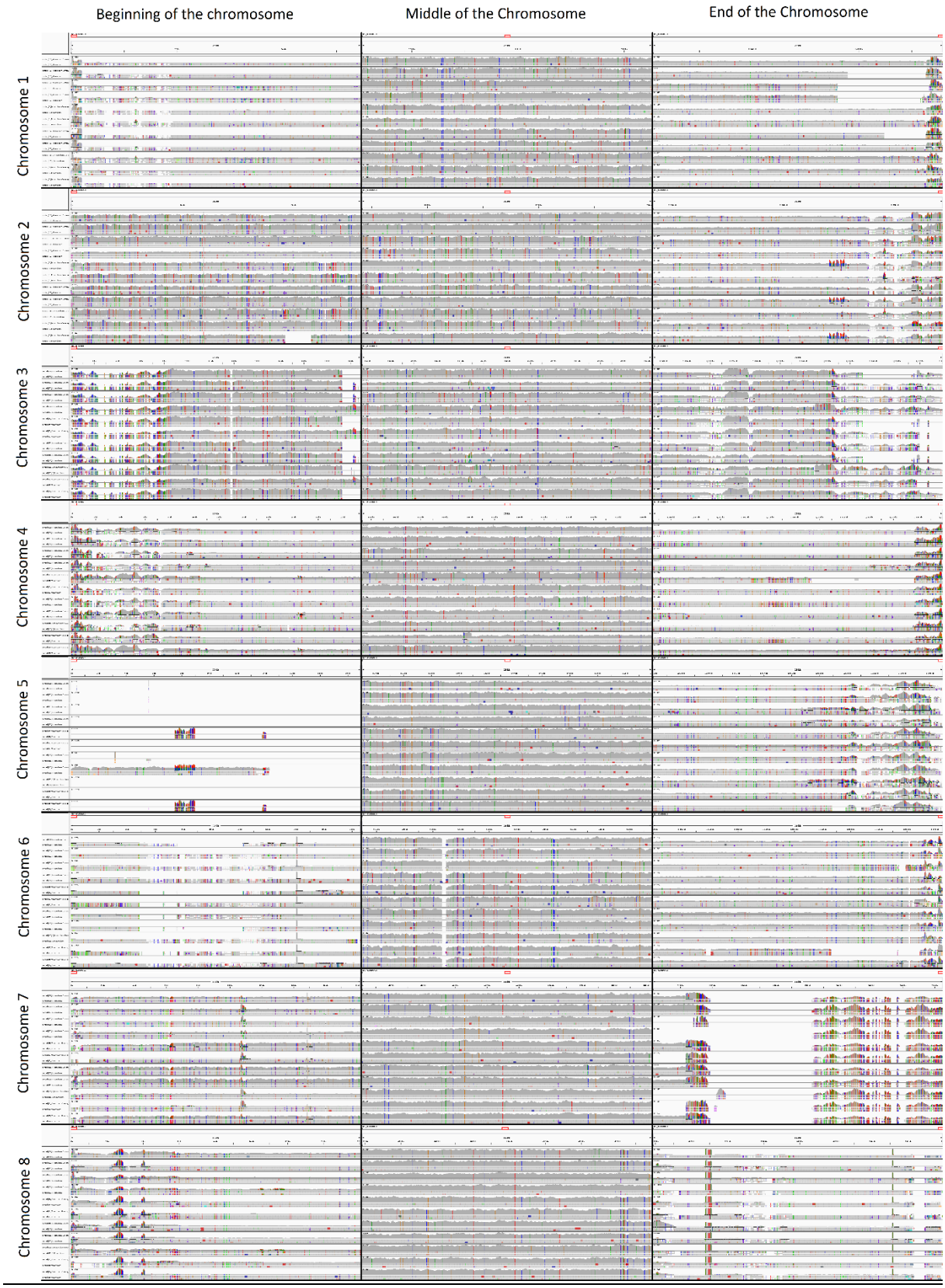


Figure 22. Visual representation of *T. delbrueckii* strains when aligned with BWA and with *T. delbrueckii CBS* 1146 as the reference genome.

8 REFERENCES

[1] P. E. McGovern *et al.*, "Fermented beverages of pre- and proto-historic China," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 101, no. 51, pp. 17593–17598, 2004.

[2] A. Fijarczyk *et al.*, "The genome sequence of the Jean-Talon strain, an archeological tetraploid beer yeast from Québec," *bioRxiv*, p. 2020.02.11.944405, 2020.

[3] V. de Québec, "No Title." [Online]. Available: http://archeologie.ville.quebec.qc.ca/medias/galeries/accueil-quebec-siege-du-gouvernement-l-ilot-des-palais-10/10_second palais_C-000360.jpg. [Accessed: 23-Jul-2020].

[4] S. Marsit, J. B. Leducq, É. Durand, A. Marchant, M. Filteau, and C. R. Landry, "Evolutionary biology through the lens of budding yeast comparative genomics," *Nat. Rev. Genet.*, vol. 18, no. 10, pp. 581–598, 2017.

[5] A. Goffeau *et al.*, "Life with 6000 genes," *Science (80-.).*, vol. 274, no. 5287, pp. 546–567, 1996.

[6] M. Kavšček, M. Stražar, T. Curk, K. Natter, and U. Petrovič, "Yeast as a cell factory: Current state and perspectives," *Microb. Cell Fact.*, vol. 14, no. 1, pp. 1–10, 2015.

[7] A. A. Duina, M. E. Miller, and J. B. Keeney, "Budding yeast for budding geneticists: A primer on the Saccharomyces cerevisiae model system," *Genetics*, vol. 197, no. 1, pp. 33–48, 2014.

[8] D. Botstein and G. R. Fink, "Yeast: An experimental organism for 21st century biology," *Genetics*, vol. 189, no. 3, pp. 695–704, 2011.

[9] S. M. Richardson *et al.*, "Design of a synthetic yeast genome," *Science (80-.).*, vol. 355, no. 6329, pp. 1040–1044, 2017.

[10] C. P. Kurtzman and J. W. Fell, *The Yeasts, A Taxonomic Study*, Fourth edi. ELSEVIER SCIENCE B.V., 1998.

[11] S. Mita, "HOW MICROBES CREATE OUR FAVORITE DELICACIES," *Torulaspora delbrueckii*, 2016. [Online]. Available: https://fermentationstations.wordpress.com/2016/09/26/candida-milleri-stiven-mita/. [Accessed: 13-Jul-2020].

[12] D. Tibor, "Handbook of Food Spoilage Yeasts, Second Edition (Contemporary Food Science): Tibor Deak: 9781420044935: Amazon.com: Books," *Handbook of food spoilage yeasts*. p. 350, 2007.

[13] L. Hellborg and J. Piškur, "Yeast diversity in the brewing industry," *Beer Heal. Dis. Prev.*, pp. 77–88, 2008.

[14] V. van Breda, N. Jolly, and J. van Wyk, "Characterisation of commercial and natural Torulaspora delbrueckii wine yeast strains," *Int. J. Food Microbiol.*, vol. 163, no. 2–3, pp. 80–88, 2013.

[15] F. Tondini, T. Lang, L. Chen, M. Herderich, and V. Jiranek, "Linking gene expression and oenological traits: Comparison between Torulaspora delbrueckii and Saccharomyces cerevisiae strains," *Int. J. Food Microbiol.*, vol. 294, no. February, pp. 42–49, 2019.

[16] M. M. C. David L. Nelson, *Lehninger principles of biochemistry*. New York, NY: W.H. Freeman, 2017.

[17] F. W. Bai, W. A. Anderson, and M. Moo-Young, "Ethanol fermentation technologies from sugar and starch feedstocks," *Biotechnol. Adv.*, vol. 26, no. 1, pp. 89–105, 2008.

[18] I. S. Pretorius, "Tailoring wine yeast for the new millennium: Novel approaches to the ancient art of winemaking," *Yeast*, vol. 16, no. 8, pp. 675–729, 2000.

[19] T. Rossignol, L. Dulau, A. Julien, and B. Blondin, "Genome-wide monitoring of wine yeast gene expression during alcoholic fermentation," *Yeast*, vol. 20, no. 16, pp. 1369–1385, 2003.

[20] R. K. Mortimer, "Evolution and variation of the yeast (Saccharomyces) genome," *Genome Res.*, vol. 10, no. 4, pp. 403–409, 2000.

[21] B. Llorente *et al.*, "Genomic Exploration of the Hemiascomycetous Yeasts: 20. Evolution of gene redundancy compared to Saccharomyces cerevisiae," *FEBS Lett.*, vol. 487, no. 1, pp. 122–133, 2000.

[22] J. H. Swiegers and I. S. Pretorius, "Yeast modulation of wine flavor," *Adv. Appl. Microbiol.*, vol. 57, no. SUPPL. A, pp. 131–175, 2005.

[23] C. Bidenne, B. Blondin, S. Dequin, and F. Vezinhet, "Analysis of the chromosomal DNA polymorphism of wine strains of Saccharomyces cerevisiae," *Curr. Genet.*, vol. 22, no. 1, pp. 1–7, 1992.

[24] A. T. BAKALINSKY and R. SNOW, “The Chromosomal Constitution of Wine Strains of Saccharomyces cerevisiae,” *YEAST*, vol. 6, pp. 367–382, 1990.

[25] N. P. Jolly, C. Varela, and I. S. Pretorius, “Not your ordinary yeast: Non-Saccharomyces yeasts in wine production uncovered,” *FEMS Yeast Res.*, vol. 14, no. 2, pp. 215–237, 2014.

[26] P. Renault, J. Coulon, G. de Revel, J. C. Barbe, and M. Bely, “Increase of fruity aroma during mixed T. delbrueckii/S. cerevisiae wine fermentation is linked to specific esters enhancement,” *Int. J. Food Microbiol.*, vol. 207, pp. 40–48, 2015.

[27] S. Benito, “The impact of Torulaspora delbrueckii yeast in winemaking,” *Appl. Microbiol. Biotechnol.*, vol. 102, no. 7, pp. 3081–3094, 2018.

[28] S. Visintin, L. Ramos, N. Batista, P. Dolci, F. Schwan, and L. Cocolin, “Impact of Saccharomyces cerevisiae and Torulaspora delbrueckii starter cultures on cocoa beans fermentation,” *Int. J. Food Microbiol.*, vol. 257, no. January, pp. 31–40, 2017.

[29] Y. Lu, J. Y. Chua, D. Huang, P. R. Lee, and S. Q. Liu, “Biotransformation of chemical constituents of durian wine with simultaneous alcoholic fermentation by Torulaspora delbrueckii and malolactic fermentation by Oenococcus oeni,” *Appl. Microbiol. Biotechnol.*, vol. 100, no. 20, pp. 8877–8888, 2016.

[30] M. Combina, A. Elía, L. Mercado, C. Catania, A. Ganga, and C. Martinez, “Dynamics of indigenous yeast populations during spontaneous fermentation of wines from Mendoza, Argentina,” *Int. J. Food Microbiol.*, vol. 99, no. 3, pp. 237–243, 2005.

[31] P. Renault *et al.*, “Genetic characterization and phenotypic variability in Torulaspora delbrueckii species: Potential applications in the wine industry,” *Int. J. Food Microbiol.*, vol. 134, no. 3, pp. 201–210, 2009.

[32] M. Lambrechts, R. Cordero-otero, and I. S. Pretorius, “Development and assessment of a recombinant Saccharomyces cerevisiae wine yeast producing two aroma-enhancing β - glucosidases encoded by the Saccharomycopsis fibuligera BGL1 and BGL2 genes,” *Ann. Microbiol.*, vol. 55, no. January, pp. 33–42, 2005.

[33] P. Hernández-Orte, M. Cersosimo, N. Loscos, J. Cacho, E. Garcia-Moruno, and V. Ferreira, “The development of varietal aroma from non-floral grapes by yeasts of different genera,” *Food Chem.*, vol. 107, no. 3, pp. 1064–1077, 2008.

[34] M. Sadoudi *et al.*, "Yeast-yeast interactions revealed by aromatic profile analysis of Sauvignon Blanc wine fermented by single or co-culture of non-Saccharomyces and Saccharomyces yeasts," *Food Microbiol.*, vol. 32, no. 2, pp. 243–253, 2012.

[35] Á. Benito, F. Calderón, and S. Benito, "The influence of non-saccharomyces species on wine fermentation quality parameters," *Fermentation*, vol. 5, no. 3, pp. 1–18, 2019.

[36] H. H. Nieuwoudt, B. A. Prior, L. S. Pretorius, and F. F. Bauer, "Glycerol in South African Table Wines: An Assessment of its Relationship to Wine Quality," *South African J. Enol. Vitic.*, vol. 23, no. 1, pp. 22–30, 2002.

[37] P. Domizio *et al.*, "Outlining a future for non-Saccharomyces yeasts: Selection of putative spoilage wine strains to be used in association with Saccharomyces cerevisiae for grape juice fermentation," *Int. J. Food Microbiol.*, vol. 147, no. 3, pp. 170–180, 2011.

[38] H. Guth, "Identification of Character Impact Odorants of Different White Wine Varieties," *J. Agric. Food Chem.*, vol. 45, no. 8, pp. 3022–3026, 1997.

[39] M. Azzolini *et al.*, "Effects of Torulaspora delbrueckii and Saccharomyces cerevisiae mixed cultures on fermentation and aroma of Amarone wine," *Eur. Food Res. Technol.*, vol. 235, no. 2, pp. 303–313, 2012.

[40] C. L. and M. J. S. Andreia Pacheco, Júlia Santos, Susana Chaves, Judite Almeida, "The Emerging Role of the Yeast Torulaspora delbrueckii in Bread and Wine Production: Using Genetic Manipulation to Study Molecular Basis of Physiological Responses," in *Structure and Function of Food Engineering*, 2012, pp. 339–370.

[41] F. Sanger, S. Nicklen, and A. . Coulson, "DNA sequencing with chain-terminating," *Proc Natl Acad Sci USA*, vol. 74, no. 12, pp. 5463–5467, 1977.

[42] M. Margulies *et al.*, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, no. 7057, pp. 376–380, 2005.

[43] O. Morozova and M. A. Marra, "Applications of next-generation sequencing technologies in functional genomics," *Genomics*, vol. 92, no. 5, pp. 255–264, 2008.

[44] M. Ronaghi, "Pyrosequencing sheds light on DNA sequencing," *Genome Res.*, vol. 11, no. 1, pp. 3–11, 2001.

[45] M. Kchouk, J. F. Gibrat, and M. Elloumi, “Generations of Sequencing Technologies: From First to Next Generation,” *Biol. Med.*, vol. 09, no. 03, 2017.

[46] J. Brind’Amour, “Flow cytometry analysis and sorting of chromosomes following hybridization with fluorescent probes that target specific DNA repeat sequences,” 2011.

[47] R. N. Bharagava, D. Purchase, G. Saxena, and S. I. Mulla, *Applications of Metagenomics in Microbial Bioremediation of Pollutants*. Elsevier Inc., 2019.

[48] J. A. Reuter, D. V. Spacek, and M. P. Snyder, “High-Throughput Sequencing Technologies,” *Mol. Cell*, vol. 58, no. 4, pp. 586–597, 2015.

[49] K. Kumar Jadav, A. Pratap Singh, A. B. Srivastav, and B. C. Sarkhel, “Molecular characterization of the complete mitochondrial genome sequence of Indian wild pig (Sus scrofa cristatus),” *Anim. Biotechnol.*, vol. 30, no. 2, pp. 186–191, 2019.

[50] I. A. rights reserved. 2020 Illumina, “Illumina.” [Online]. Available: https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html. [Accessed: 11-Oct-2021].

[51] E. C. Hayden, “Is the $1,000 genome for real?,” *January 15*, 2014. [Online]. Available: https://www.nature.com/news/is-the-1-000-genome-for-real-1.14530.

[52] R. Bao *et al.*, “Review of current methods, Applications, And data management for the bioinformatics analysis of whole exome sequencing,” *Cancer Inform.*, vol. 13, pp. 67–82, 2014.

[53] B. Institute, “Babraham Bioinformatics,” *FastQC*. [Online]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. [Accessed: 13-Sep-2020].

[54] S. Andrews, “FastQC 1 . 1 What is FastQC 2 . Basic Operations 2 . 1 Opening a Sequence file,” 2010.

[55] J. Il Sohn and J. W. Nam, “The present and future of de novo whole-genome assembly,” *Brief. Bioinform.*, vol. 19, no. 1, pp. 23–40, 2018.

[56] A. M. Giani, G. R. Gallo, L. Gianfranceschi, and G. Formenti, “Long walk to genomics: History and current approaches to genome sequencing and assembly,” *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 9–19, 2020.

[57] D. R. Zerbino and E. Birney, “Velvet: Algorithms for de novo short read assembly using de Bruijn

graphs," *Genome Res.*, vol. 18, no. 5, pp. 821–829, 2008.

[58] S. D. Jackman *et al.*, "ABySS 2 . 0 : Resource-Efficient Assembly of Large Genomes using a Bloom Filter Effect of Bloom Filter False Positive Rate," *Genome Res.*, vol. 27, pp. 768–777, 2017.

[59] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 26, no. 5, pp. 589–595, 2010.

[60] B. Langmead, "Aligning short sequencing reads with Bowtie," *Curr. Protoc. Bioinforma.*, no. SUPP.32, pp. 1–14, 2010.

[61] A. Bankevich *et al.*, "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing," *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, 2012.

[62] R. Li *et al.*, "SOAP2: An improved ultrafast tool for short read alignment," *Bioinformatics*, vol. 25, no. 15, pp. 1966–1967, 2009.

[63] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[64] "Burrows-Wheeler Aligner." [Online]. Available: http://bio-bwa.sourceforge.net/. [Accessed: 07-Oct-2020].

[65] G. R. Limited, "Samtools," 2019. [Online]. Available: http://www.htslib.org/.

[66] H. Thorvaldsdóttir, J. T. Robinson, and J. P. Mesirov, "Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration," *Brief. Bioinform.*, vol. 14, no. 2, pp. 178–192, 2013.

[67] I. Milne *et al.*, "Tablet-next generation sequence assembly visualization," *Bioinformatics*, vol. 26, no. 3, pp. 401–402, 2009.

[68] T. Carver, S. R. Harris, T. D. Otto, M. Berriman, J. Parkhill, and J. A. McQuillan, "BamView: Visualizing and interpretation of next-generation sequencing read alignments," *Brief. Bioinform.*, vol. 14, no. 2, pp. 203–212, 2013.

[69] M. Fiume, V. Williams, A. Brook, and M. Brudno, "Savant: Genome browser for high-throughput sequencing data," *Bioinformatics*, vol. 26, no. 16, pp. 1938–1944, 2010.

[70] T. Carver, S. R. Harris, M. Berriman, J. Parkhill, and J. A. McQuillan, “Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data,” *Bioinformatics*, vol. 28, no. 4, pp. 464–469, 2012.

[71] J. T. Robinson, H. Thorvaldsdóttir, A. M. Wenger, A. Zehir, and J. P. Mesirov, “Variant review with the integrative genomics viewer,” *Cancer Res.*, vol. 77, no. 21, pp. e31–e34, 2017.

[72] V. Dominguez Del Angel *et al.*, “Ten steps to get started in Genome Assembly and Annotation,” *F1000Research*, vol. 7, p. 148, 2018.

[73] E. Proux-Wéra, D. Armisén, K. P. Byrne, and K. H. Wolfe, “A pipeline for automated annotation of yeast genome sequences by a conserved-synteny approach,” *BMC Bioinformatics*, vol. 13, no. 1, 2012.

[74] J. M. Mudge and J. Harrow, “The state of play in higher eukaryote gene annotation,” *Nat. Rev. Genet.*, vol. 17, no. 12, pp. 758–772, 2016.

[75] J. Huerta-Cepas *et al.*, “Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper,” *Mol. Biol. Evol.*, vol. 34, no. 8, pp. 2115–2122, 2017.

[76] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “Data, information, knowledge and principle: Back to metabolism in KEGG,” *Nucleic Acids Res.*, vol. 42, no. D1, pp. 199–205, 2014.

[77] S. Choudhuri and S. Choudhuri, “Chapter 9 – Phylogenetic Analysis,” *Bioinforma. Beginners*, pp. 209–218, 2014.

[78] P. Kapli, Z. Yang, and M. J. Telford, “Phylogenetic tree building in the genomic age,” *Nat. Rev. Genet.*, vol. 21, no. 7, pp. 428–444, 2020.

[79] C. Peng, “Distance based methods in phylogenetic tree construction,” *Neural, Parallel Sci. Comput.*, vol. 15, no. 4, pp. 547–560, 2007.

[80] Z. Yang, *Computational Molecular Evolution*. 2006.

[81] M. Y. Galperin, K. S. Makarova, Y. I. Wolf, and E. V. Koonin, “Expanded Microbial genome coverage and improved protein family annotation in the COG database,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D261–D269, 2015.

[82] F. Tondini, V. Jiranek, P. R. Grbin, and C. A. Onettoa, “Genome Sequence of Australian

Indigenous Wine Yeast Torulaspora Delbrueckii COFT1 Using Nanopore Sequencing.," *Genome Announc.*, pp. 1–2, 2018.

[83] F. Sievers and D. G. Higgins, "Clustal Omega for making accurate alignments of many protein sequences," *Protein Sci.*, vol. 27, no. 1, pp. 135–145, 2018.

[84] L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh, "IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies," *Mol. Biol. Evol.*, vol. 32, no. 1, pp. 268–274, 2015.

[85] "samtools-sort." [Online]. Available: http://www.htslib.org/doc/samtools-sort.html. [Accessed: 05-Jul-2020].

[86] "samtools-index." [Online]. Available: http://www.htslib.org/doc/samtools-index.html. [Accessed: 05-Jul-2020].

[87] "samtools-view." [Online]. Available: http://www.htslib.org/doc/samtools-view.html. [Accessed: 05-Jul-2020].

[88] "Integrative Genomics Viewer." [Online]. Available: http://software.broadinstitute.org/software/igv/. [Accessed: 19-Jan-2021].

[89] B. Institute, and the R. of The, and U. of California, "Integrative Genomics Viewer." [Online]. Available: https://software.broadinstitute.org/software/igv/LoadGenome.

[90] M. G. Garrison E, "Haplotype-based variant detection from short-read sequencing," *arXiv Prepr. arXiv1207.3907 [q-bio.GN]*, 2012.

[91] "Freebayes." [Online]. Available: https://github.com/freebayes/freebayes. [Accessed: 19-Jan-2021].

[92] "samtools flagstat." [Online]. Available: http://www.htslib.org/doc/samtools-flagstat.html. [Accessed: 19-Jan-2021].

[93] "vcf-stats." [Online]. Available: https://vcftools.github.io/perl_module.html#vcf-stats. [Accessed: 19-Jan-2021].

[94] "spades." [Online]. Available: https://github.com/ablab/spades. [Accessed: 20-Jan-2021].

[95] H. Li, "Minimap2: Pairwise alignment for nucleotide sequences," *Bioinformatics*, vol. 34, no. 18,

pp. 3094–3100, 2018.

[96] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[97] H. Li, “A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data,” *Bioinformatics*, vol. 27, no. 21, pp. 2987–2993, 2011.

[98] “SequenceTools - pileupCaller.” [Online]. Available: https://github.com/stschiff/sequenceTools. [Accessed: 07-Jan-2021].

[99] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, “Principal components analysis corrects for stratification in genome-wide association studies,” *Nat. Genet.*, vol. 38, no. 8, pp. 904–909, 2006.

[100] J. L. Gordon, D. Armisen, E. Proux-Weŕa, S. S. OhÉigeartaigh, K. P. Byrne, and K. H. Wolfe, “Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 50, pp. 20024–20029, 2011.

[101] J. Gomez-Angulo *et al.*, “Genome sequence of Torulaspora delbrueckii NRRL Y-50541, isolated from mezcal fermentation,” *Genome Announc.*, vol. 3, no. 4, pp. 3–4, 2015.

[102] R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin, “The COG database: A tool for genome-scale analysis of protein functions and evolution,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 33–36, 2000.

[103] T. Gabaldón and E. V. Koonin, “Functional and evolutionary implications of gene orthology,” *Nat. Rev. Genet.*, vol. 14, no. 5, pp. 360–366, 2013.

[104] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Res.*, vol. 28, pp. 27–30, 2020.

[105] D. W. K. Toh, J. Y. Chua, Y. Lu, and S. Q. Liu, “Evaluation of the potential of commercial non-Saccharomyces yeast strains of Torulaspora delbrueckii and Lachancea thermotolerans in beer fermentation,” *Int. J. Food Sci. Technol.*, vol. 55, no. 5, pp. 2049–2059, 2020.

[106] J. Molinet, F. A. Cubillos, F. Salinas, G. Liti, and C. Martínez, “Genetic variants of TORC1 signaling pathway affect nitrogen consumption in Saccharomyces cerevisiae during alcoholic

fermentation," *PLoS One*, vol. 14, no. 7, pp. 1–24, 2019.

[107] J. A. Diderich *et al.*, "Glucose uptake kinetics and transcription of HXT genes in chemostat cultures of Saccharomyces cerevisiae," *J. Biol. Chem.*, vol. 274, no. 22, pp. 15350–15359, 1999.

[108] L. Ye, A. L. Kruckeberg, J. A. Berden, and K. Van Dam, "Growth and glucose repression are controlled by glucose transport in Saccharomyces cerevisiae cells containing only one glucose transporter," *J. Bacteriol.*, vol. 181, no. 15, pp. 4673–4675, 1999.

[109] C. Alves-Araújo, M. J. Hernandez-Lopez, J. A. Prieto, F. Randez-Gil, and M. J. Sousa, "Isolation and characterization of the LGT1 gene encoding a low-affinity glucose transporter from Torulaspora delbrueckii," *Yeast*, vol. 22, no. 3, pp. 165–175, 2005.

[110] M. J. S. Andreia Pacheco, Lorena Donzella, Maria Jose Hernandez-Lopez, Maria Judite Almeida, Jose Antonio Prieto, Francisca Randez-Gil, John P Morrissey, "Hexose Transport in Torulaspora Delbrueckii: Identification of Igt1, a New Dual-Affinity Transporter.," *FEMS Yeast Res.*, vol. 20, no. 1, 2020.

[111] L. Petruzzi *et al.*, "Microbial resources and enological significance: Opportunities and benefits," *Front. Microbiol.*, vol. 8, no. JUN, pp. 1–13, 2017.

[112] C. Nadai, L. Treu, S. Campanaro, A. Giacomini, and V. Corich, "Different mechanisms of resistance modulate sulfite tolerance in wine yeasts," *Appl. Microbiol. Biotechnol.*, vol. 100, no. 2, pp. 797–813, 2016.

[113] R. Franco-Duarte *et al.*, "New integrative computational approaches unveil the Saccharomyces cerevisiae pheno-metabolomic fermentative profile and allow strain selection for winemaking," *Food Chem.*, vol. 211, pp. 509–520, 2016.

[114] D. Franco-Duarte, R.; Mendes, I.; Umek, L.; Drumonde-Neves, J.; Zupan, B.; Schuller, "Computational Models Reveal Genotype-Phenotype Associations in Saccharomyces cerevisiae.," *Yeast*, no. 31, pp. 265–277, 2014.

[115] S. Wong and K. H. Wolfe, "Duplication of genes and genomes in yeasts," *Top. Curr. Genet.*, vol. 15, no. January, pp. 79–99, 2006.

[116] K. H. Wolfe, "Origin of the yeast whole-genome duplication," *PLoS Biol.*, vol. 13, no. 8, pp. 1–7, 2015.

[117] I. . James, S.A.; Collins, M.D.; Roberts, "Use of an RRNA Internal Transcribed Spacer Region to Distinguish Phylogenetically Closely Related Species of the Genera Zygosaccharomyces and Torulaspora," *Int. J. Syst. Bacteriol.*, vol. 46, pp. 189–194, 1996.

[118] E. Aller-Arranz, F. Randez-Gil, E. Barrio, and J. A. Prieto, "A DNA region of Torulaspora delbrueckii containing the HIS3 gene: Sequence, gene order and evolution," *Yeast*, vol. 20, no. 16, pp. 1359–1368, 2003.

[119] R. Escribano *et al.*, "Wine aromatic compound production and fermentative behaviour within different non-Saccharomyces species and clones," *J. Appl. Microbiol.*, vol. 124, no. 6, pp. 1521–1531, 2018.

[120] C. P. Kurtzman and C. J. Robnett, "Phylogenetic relationships among yeasts of the 'Saccharomyces complex' determined from multigene sequence analyses," *FEMS Yeast Res.*, vol. 3, no. 4, pp. 417–432, 2003.

[121] C. P. Kurtzman, "Phylogenetic circumscription of Saccharomyces, Kluyveromyces and other members of the Saccharomycetaceae, and the proposal of the new genera Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma and Zygotorulaspora," *FEMS Yeast Res.*, vol. 4, no. 3, pp. 233–245, 2003.

[122] M. Stratford, "Food and Beverage Spoilage Yeasts," in *The Yeast Handbook*, G. H. Querol, Amparo, Fleet, Ed. Springer-Verlag: Berlin/Heidelberg, 2006, pp. 335–377.

[123] P. Martorell, M. Stratford, H. Steels, M. T. Fernández-Espinar, and A. Querol, "Physiological characterization of spoilage strains of Zygosaccharomyces bailii and Zygosaccharomyces rouxii isolated from high sugar environments," *Int. J. Food Microbiol.*, vol. 114, no. 2, pp. 234–242, 2007.

[124] D. P. and P. B. Nurzhan Kuanyshev , Giusy M. Adamo, "The spoilage yeast Zygosaccharomyces bailii: Foe or friend?," *Yeast*, pp. 359–370, 2017.

[125] A. Vilela, "Use of nonconventional yeasts for modulating wine acidity," *Fermentation*, vol. 5, no. 1, 2019.

[126] M. Palma, J. F. Guerreiro, and I. Sá-Correia, "Adaptive response and tolerance to acetic acid in Saccharomyces cerevisiae and Zygosaccharomyces bailii: A physiological genomics perspective," *Front. Microbiol.*, vol. 9, no. FEB, pp. 1–16, 2018.

[127] M. Palma and I. Sá-Correia, *Physiological Genomics of the Highly Weak-Acid-Tolerant Food Spoilage Yeasts of Zygosaccharomyces bailii sensu lato*, vol. 58. Springer International Publishing, 2019.

[128] H. Guo *et al.*, "Genomic Insights Into Sugar Adaptation in an Extremophile Yeast Zygosaccharomyces rouxii," *Front. Microbiol.*, vol. 10, no. February, pp. 1–10, 2020.

[129] N. P. Mira *et al.*, "The genome sequence of the highly acetic acid-tolerant zygosaccharomyces bailii-derived interspecies hybrid strain ISA1307, isolated from a sparkling wine plant," *DNA Res.*, vol. 21, no. 3, pp. 299–313, 2014.

[130] J. Santos *et al.*, "Ethanol tolerance of sugar transport, and the rectification of stuck wine fermentations," *Microbiology*, vol. 154, no. 2, pp. 422–430, 2008.

[131] P. Domizio, J. F. House, C. M. L. Joseph, L. F. Bisson, and C. W. Bamforth, "Lachancea thermotolerans as an alternative yeast for the production of beer," *J. Inst. Brew.*, vol. 122, no. 4, pp. 599–604, 2016.

[132] Á. Benito, F. Calderón, F. Palomero, and S. Benito, "Quality and composition of airén wines fermented by sequential inoculation of lachancea thermotolerans and saccharomyces cerevisiae," *Food Technol. Biotechnol.*, vol. 54, no. 2, pp. 135–144, 2016.

[133] K. Thompson Witrick, S. Duncan, K. Hurley, and S. O'Keefe, "Acid and Volatiles of Commercially-Available Lambic Beers," *Beverages*, vol. 3, no. 4, p. 51, 2017.

[134] X. X. Shen *et al.*, "Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum," *Cell*, vol. 175, no. 6, pp. 1533-1545.e20, 2018.

[135] A. Roetzer, T. Gabaldón, and C. Schüller, "From Saccharomyces cerevisiae to Candida glabrata in a few easy steps: Important adaptations for an opportunistic pathogen," *FEMS Microbiol. Lett.*, vol. 314, no. 1, pp. 1–9, 2011.

[136] A. Y. Coughlan *et al.*, "The yeast mating-type switching endonuclease HO is a domesticated member of an unorthodox homing genetic element family," *bioRxiv*, pp. 1–24, 2020.