Deciphering the *Debaryomyces* spp. complex using Next Generation Sequencing: Are they the same species?

By

Ramón E. Rivera-Vicéns

A thesis submitted in partial fulfillment of the requirements for the degree of
MASTER IN SCIENCE
In
MARINE SCIENCES
BIOLOGICAL OCEANOGRAPHY
UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2018

Approved by:

_____	_____
Govind S. Nadathur, Ph.D. President, Graduate Committee	Date
_____	_____
Nikolaos Schizas, Ph.D. Member, Graduate Committee	Date
_____	_____
Lorenzo Saliceti, Ph.D. Member, Graduate Committee	Date
_____	_____
Lillian Ramirez, M.S. Graduate Studies Representative	Date
_____	_____
Ernesto Otero Morales, Ph.D. Director of Department	Date

ABSTRACT

Debaryomyces hansenii is a unicellular marine yeast, from the Phylum Ascomycota and is widely distributed in nature. It has been isolated from high salt environments like estuaries and bays, to salty foods and cheese. This yeast has a rounded morphology, is non-mobile and can withstand up to 12% of NaCl (w/v) and pH levels between 3.0-10.0 (optimal pH=7). Some strains can even withstand up to 25% of NaCl (w/v), making these strains extreme halophiles. The current classification of this yeast differentiates *Debaryomyces* spp. into two independent taxonomic entities, *D. hansenii* and *D. fabryi*. Although many studies have been performed for the taxonomic assignment of the *Debaryomyces* genera (*i.e.* single gene sequencing, molecular characterization), most of the results in these studies have been ambiguous and still there is not a clear pattern on how to classify these organisms at the species and strain level. In this study, we employed Next Generation Sequencing (NGS) and used different bioinformatics tools for the creation of six genomes of the *Debaryoymyces* spp. These new genomes coupled with other current available genomes were used for the generation of phylogenies to aim to resolved the taxonomic assignment problem in this group of closely related species. We have showed that the use of NGS and bioinformatics can be used to taxonomically assign yeast species at the strain level. We also present data that shows how the flavinogenic capability, that some of these strains posses, could be a result of a speciation event in which divided species that are flavinogenic versus non-flavinogenic. This result was observed in our different phylogeny analysis (*i.e.* Bayesian and Maximum Likelihood) throughout the study.

RESUMEN

Debaryomyces hansenii es una levadura marina unicelular, perteneciente al Filo de los Ascomicetos, distribuida a través de la naturaleza. Ha sido aislada de ambientes con altas concentraciones de sal como estuarios y bahías, hasta en comidas saladas y quesos. Esta levadura posee una morfología circular, es no motil y puede aguantar hasta 12% de NaCl (w/v) y niveles de pH entre 3.0-10.0 (pH óptimo de 7.0). Algunas de las cepas pueden resistir hasta 25% de NaCl (w/v), por lo que son clasificados como halófilas extremas. La clasificación actual de esta levadura divide las especies de *Debaryomyces* spp. en dos entidades taxonómicas distintas: *Debaryomyces hansenii y Debaryomyces fabryi*. Aunque se han realizado varios estudios para la clasificación taxonómica de estas levaduras, aun no se tiene un amplio conocimiento de cómo clasificar este grupo de organismo al nivel de especie y cepas. En este estudio aplicamos la Secuenciación de Nueva Generación (SNG) y usamos varias herramientas de la bioinformática para la creación de seis genomas de *Debaryoymyces* spp. Estos nuevos genomas junto a data disponible públicamente fueron usados para crear filogenias las cuales resolvieran el problema de asignación taxonómica de este grupo de especies bien relacionadas entre si. Hemos demostrado como el uso de SNG y la bioinformática pueden ser usados para asignar correctamente niveles taxonómicos de levaduras al nivel de cepas. También presentamos data que demuestra como la capacidad flavinogénica, la cual alguna de estas especies poseen, puede ser resultado de un evento de especiación el cual separó las levaduras que poseen esta característica versus las que no la poseen. Este hallazgo fue obtenido en nuestros varios análisis de filogenia (Bayesiana y Probabilidad Máxima) de nuestro estudio.

COPYRIGHT

May 2018

DEDICATION

To my family.

ACKNOWLEDGEMENTS

Thanks to my committee for their constant support and invaluable lessons. To Govind, for giving me the opportunity of being part of his lab and all his patience when I needed to change my thesis project (twice) to a complete new one. Thanks also to Schizas, the first professor that let me performed research as an undergraduate student, in his lab on Magueyes, with the student (now Ph.D.) Matt Lucas. Thanks to Lorenzo, for his constant support, help and summer assistantship.

Also, I am very thankful to my lab colleagues: Mairim, Nicolle, Yariela (M.S.) and Camille. But also to Bill Rosado for the stories, fruits and fish we ate in the lunch breaks. Thanks for being always there and for all lessons you taught me as a lab technician and as a friend.

To my second lab and colleagues that always welcome me and all the collaborations we have done: Nick Hammerman, Ingrid, Alex, Liajay, Jaaziel.

Special thanks to Professor Alex Van Dam (Biology - UPRM) for letting me use his account of the Greenfield system at the Pittsburgh Supercomputing Center (PSC), under the project Extreme Science and Engineering Discovery Environment (XSEDE) (National Science Foundation grant number ACI-1548562) for generating the assemblies of the genomes.

Thanks to Ciris Energy Inc, especially Dr. Marthah Delome for her help in sequencing.

Thanks to Sea Grant College Program at UPRM for their support in materials and help for assisting conferences and workshops.

Thanks to also the Weyerhaeuser Corporation for their support and lab equipment we used for the analysis.

And very special thanks to my Family! I love you guys! All these was thanks to your constant support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1 *Debaryomyces hansenii*

Debaryomyces hansenii is a unicellular marine yeast, from the Phylum Ascomycota and is widely distributed in nature. It has been isolated from high salt environments like estuaries and bays, to salty foods and cheese (Adler *et al.*, 1986, Gadd & Edwards, 1986). Strains of *D. hansenii* have been isolated from depths of 50-100m more than 15 km off the coast of South Baja California, Mexico (Centro de Investigaciones Biológicas del Noroeste, CIBNOR). Some strains of *D. hansenii* have been isolated even from humans (Breuer and Harms, 2006, Wong *et al.*, 1982). It can use many carbon substrates like: glucose, glycerol, galactose, maltose, trehalose and raffinose, to name a few. It also can use inorganic ammonium as the only nitrogen source (Yanai *et al.*, 1994). Its fermentative capacity is very low, mainly due to the low activity of the phosphofructokinase enzyme (Sanchez *et al.*, 2006).

This yeast has a rounded morphology, is non-mobile and can withstand up to 12% of NaCl (w/v) and pH levels between 3.0-10.0 (optimal pH=7). However, some strains can withstand up to 25% of NaCl (w/v), making these strains extreme halophiles (Breuer and Harms, 2006; Butinar *et al.*, 2005). Optimal temperature of growth is approximately 20-25^{o}C (Breuer and Harms, 2006), though it can grow from 15-40^{o}C (Nguyen *et al.*, 2009). Experiments in our laboratory have shown that some strains can grow at 4^{o}C, resulting in higher cell densities than cultures growing at 25^{o}C (unpublished results). These results are in concordance with reports of isolation of *D. hansenii* strains from arctic glaciers, glacial melt-water, and seawater adjacent to the glaciers (Jacques *et al.*, 2015; Butinar *et al.*, 2011, Davenport, 1980). Furthermore, this yeast can resist high levels of pressure. When exposed to simulated deep-sea conditions, *D. hansenii* could grow up to a pressure of 20 MPa (~2000m) (Lorenz and Molitoris, 1997).

D. hansenii is found in such environments due to its high osmotolerance since *D. hansenii* is capable of accumulating high concentrations of sodium ions without affecting its biochemistry (Prista *et al.*, 1997). When *D. hansenii* is grown in the presence of high NaCl concentrations, it tends to accumulate high concentration of sodium (Na^{+}) inside

the cell (Norkrans and Kylin, 1969; Thomé-Ortiz *et al.*, 1998). Consequently, it starts sequestering sodium ions into the cell vacuole (Gonzalez-Hernandez *et al.*, 2004) and can extrude an excess of sodium (Ruiz and Ariño, 2007; Gonzalez *et al.*, 2009), resisting the toxic effect of the cations. Also, a couple of studies have shown that sodium plays a significant role in the growth of *D. hansenii* at environments with additional stress factors like high temperature and serious pH levels (*e.g.* 3.5, 7.8) (Almagro *et al.*, 2000; Gonzalez- Hernandez *et al.*, 2004; Prista *et al.*, 2005). In fact, when *D. hansenii* is in oxidative stress, NaCl exert a protective effect in this yeast (Navarrete *et al.*, 2009). The genes encoding Na^+ ATPases (*i.e.* ENA1 and ENA2) (Ruiz and Ariño, 2007) from *D. hansenii* have been successfully cloned in a mutant of *S. cerevisiae* lacking the sodium efflux systems and tolerance to sodium (Almagro *et al.*, 2001). After cloning, this yeast had the capability to tolerate and extrude sodium.

D. hansenii possess a majority of halotolerance genes (*i.e.* genes in charge of improving salt tolerance) that were described for the model yeast *S. cerevisiae* (*e.g.* HAL2, HAL3, HAL4), with an exception of HAL1 (Gaxiola *et al.*, 1992; Rios *et al.*, 1997; Ferrando *et al.*, 1995). Some of these genes of *D. hansenii* (*i.e.* HAL2) have been successfully isolated and characterized (Aggarwal *et al.*, 2005). The protein encoded by the gene DHAL2 (*i.e.* Dhal2p), showed significantly higher resistance towards lithium and sodium ions in comparison with protein homologues from other yeast (Aggarwal *et al.*, 2005). Moreover, *D. hansenii* produces two distinct isoforms of the enzyme encoded by DHAL2 (Aggarwal and Mondal, 2006); one is cytosolic and the other is membrane bound. The study also demonstrated that the membrane bound isoform of Dhal2p is unique in *D. hansenii* compared to other yeast, and it played an important role under high salt stress conditions (Aggarwal and Mondal, 2006).

Another distinctive trait of *D. hansenii* is the translation of the CTG codon as a serine amino acid instead of leucine (Massey *et al.*, 2003). This trait was first discovered in the yeast *Candida cylindracea* using cell-free translation experiments (Kawaguchi *et al.*, 1989; Ohama *et al.*, 1993). Interestingly, the tRNAs responsible for decoding serine instead of leucine, have been shown to be widely shared in yeast species from the genus *Candida*: *Candida lusitaniae, Candida guilliermondii, Candida albicans, Candida*

dubliniensis, Candida tropicalis, Candida parapsilosis and many others (Ueda *et al.*, 1994). This trait has also been used in phylogenetic studies between species capable of this special translation (Sugita *et al.*, 1999; Fitzpatrick *et al.*, 2006), and researchers now refer to this group as the CTG (or CUG) group or clade.

The type strain of *D. hansenii* is CBS 767 and its genome has been sequenced (Dujon *et al.*, 2004). It possesses a haploid genome with seven chromosomes (ranging in size from 1.25 Mb to 2.33 Mb), where most genes are single copy. This lack of complexity offers a great advantage when compared to other yeast, where annotation problems arise because of the many copies of the genes in diploid genomes (Rozpędowska *et al.*, 2011). *D. hansenii* genome size is 12.2 million bases (Mb) with a GC content of 36.3%, and encodes for 6,906 genes. Horizontal gene transfer, a phenomenon that is rare in hemiascomycetes yeast, has also been detected in *D. hansenii* (Dujon *et al.*, 2004). Researchers have also shown the presence of several linear plasmids in *D. hansenii* (*i.e.* pDHL1 (8.4 kb), pDHL2 (9.2 kb) and pDHL3 (15.0 kb); Gunge *et al.*, 1993) (pDH1A and pDH1B; Cong *et al.*, 1994).

1.2 Taxonomy and Systematics

- Current *Debaryomyces hansenii* taxonomic classification

(*Specifies a common classification and not a real taxonomic rank or division)

Super Kingdom: Eukaryota;
*Opisthokonta;
Kingdom: Fungi;
Sub Kingdom: Dikarya;
Phylum: Ascomycota;
*Saccharomyceta;
Sub Phylum: Saccharomycotina;
Class: Saccharomycetes;
Order: Saccharomycetales;
Family: Debaryomycetaceae[1];
Genus: *Debaryomyces*;

Species: *hansenii*

[1]Debaryomycetaceae family established by Kurtzman and Suzuki (2010).

Based on DNA hybridization, a group of researchers investigated various strains of *D. hansenii* and divided them into two categories, *D. hansenii* var. *hansenii* (anamorph *Candida famata* var. *famata*) and *D. hansenii* var. *fabryi* (anamorph *Candida famata* var. *flareri*) (Nakase and Suzuki, 1985a, b). A study of the 18S sequences of different yeasts resulted in placing *D. hansenii* within the hemiascomycetes subdivision (Wilmotte *et al.*, 1993, Prillinger *et al.* 1999). Nakase and Suzuki divided *D. hansenii var. hansenii* and *D. hansenii var. fabryi*, using new tools such as RAPD-PCR (Random Amplification of Polymorphic DNA Polymerase Chain Reaction). Later, Corredor *et al.* (2000) confirmed this classification using Southern Hybridization techniques with specific probes for species of *Debaryomyces*. The same group of researchers showed that species of *D. hansenii* possess high levels of chromosomal length polymorphism using a Pulsed field gel electrophoresis (PFGE) technique (Corredor *et al.* 2003). Another important trait for species differentiation is riboflavin production, where only *D. fabryi* specimens tend to be producers and *D. hansenii* are not (Nguyen *et al.*, 2009, Breuer & Harms, 2006). Although, it should be noted that there are some strains classified as *D. hansenii* with the capability of producing riboflavin. Temperature related growth have also been used to identify strains, since *D. hansenii* is characterized as having optimal growth at lower temperatures (31-35°C), compared to *D. fabryi* (36- 40°C) (Groenewald *et al.*, 2008; Nguyen *et al.*, 2009). However, some species of *D. hansenii* can grow up to 40°C (Nguyen *et al.*, 2009).

Gene sequence analysis of conserved genes (*e.g.* actin: ACT1, glycerol-3-phosphate dehydrogenase: GPD1), intergenic spacer of rRNA (*i.e.* ITS) and microsatellites have also been commonly used for strain differentiation (Desnos- Oliver *et al.*, 2008, Groenewald *et al.*, 2008, Nguyen *et al.*, 2009, Gallardo *et al.*, 2014, Lopandic *et al.*, 2013). Other researchers, employed tools like Fourier-transform infrared (FT-IR) micro-spectroscopy to identify yeast species, including *D. hansenii* (Wenning *et al.*, 2002). Another interesting approach was developed by de Silóniz *et al.*, (2000), where they

made use of a selective and differential culture medium to detect common yeast frequently associated in the spoilage of some intermediate moisture foods (*i.e.* foods with low water activity a_w). In their methodology, they used chromogenic substrates salmon-Gluc and X-Gal to detect two enzymes, β-glucosidase and β-galactosidase, obtaining reliable results for detecting *D. hansenii* from over 140 species of yeast and bacteria.

The current classification of this yeast, applying new molecular techniques, follow the results obtained by Prillinger *et al.* (1999) and other researchers. This yeast was differentiated into two independent taxonomic entities, *D. hansenii* and *D. fabryi* (Nguyen *et al.*, 2009). Since *D. hansenii* is a member of a closely related group of species, there may have been taxonomic misidentification in this yeast, since separation of species using phenotypic essays is difficult and results are very similar (Suzuki *et al.*, 2011). In fact, reports by Phaff *et al.* showed that some species believed to be *D. hansenii*, due to phenotypic traits, belonged to a new species *Debaryomyces prosopidis* (1998). In conclusion, most of the results in these studies have been ambiguous and still there is not a clear pattern on how we can classify these organisms at the species and strain level.

1.3 Phylogenies: from single genes to the genomic era

Phylogenies for yeast species have undergo a significant transformation, starting with single genes to now full genomes. First, it started by data generated by polymerase chain reaction (PCR) (James *et al.* 2006, Kurtzman and Robnett 2003). For example, in a study by Kurtzman and Robnett, they resolved phylogenies using sequences of 26S rDNA from approximately 500 species of ascomycetes yeasts (1998). By 2003, the same group of researchers generated phylogenies for 75 species using many sequences like: rDNA regions (*i.e.*18S, 26S, ITS), single copy nuclear genes (*e.g.* actin-1, RNA polymerase II) and mitochondrial genes (*e.g.* cytochrome oxidase II) (Kurtzman and Robnett, 2003). This methodology has presented sound results to resolve phylogenies but it is lab and time intensive.

The ease and availability of new genomes thanks to the application of Next Generation Sequencing (NGS), have made phylogenetic analysis far more powerful, since inferences are now made based on a much larger number of genes (Ciccarelli *et al.* 2006). NGS allows for millions of reads to be sequenced for the construction of draft genomes and the development of genes prediction and annotation. It is a comparatively cheaper technology for deep level phylogenies, thus it is monetarily accessible. NGS has been used successfully to construct phylogenies in various organisms. In the case of fungi, one research group took currently available genomes and generated phylogenies using sequences from 17 fungal species. In their methodology, they searched the orthologs proteins (*i.e.* shared genes in different species that are from the same ancestor) between them and managed to create phylogenies using a concatenation of 781 orthologs sequences (Robbertse *et al.*, 2006). Another investigation used conserved eukaryotic orthologs sequences to reconstruct phylogenies among 21 fungal genomes, three animal genomes, and one plant genome (Kuramae *et al.*, 2006). Their analysis showed a single tree with high support values for the 531 orthologs sequences after using different methods for constructing the phylogenies.

Though there have been a couple of studies addressing the classification of yeast using NGS, they have been concerned with different taxonomic Classes, Orders and Families (Shen *et al.*, 2016, Fitzpatrick *et al.*, 2006). Only a couple of studies have worked with members of the same family (Fitzpatrick *et al.*, 2006; Hansen *et al.*, 2013) but not all of them applied NGS to generate phylogenies. To our knowledge, no other studies have applied NGS to resolve phylogenies of members of the same species and/or strains. In the case of the genera *Debaryomyces*, some NGS projects have been completed in the last years and genomes have been created. Most of these genomes are from the species *hansenii* (*D. hansenii* CBS 767, Dujon *et al.*, 2004; *D. hansenii* MTCC 234, Kumar *et al.*, 2012; *D. hansenii* J6, Berrocal *et al.*, 2016) and one from the species *fabryi* (*D. fabryi* CBS 789; Tafer *et al.*, 2016). None of these genomes have been used for the generation of phylogenies except for *D. hansenii* CBS 767 (Dujon *et al.*, 2004, Shen *et al.*, 2016, Fitzpatrick *et al.*, 2006).

A recent study made use of all the existing genomes of *Debaryomyces* spp. and made a brief comparison, where they mapped the sequencing reads of *D. hansenii* strain J6 with the other available genomes (Berrocal *et al.*, 2016). Overall results of the mapping showed that MTCC 234 had the highest similarity to J6, followed by CBS 789 and CBS 767. This was an unexpected result since CBS 789 is classified as *D. fabryi* and not as *D. hansenii*. Our laboratory performed protein sequence similarity searches using a UniProt custom database containing all the proteins of the Debaryomycetaceae family (Kurtzman and Suzuki, 2010) with the program MAKER (Cantarel *et al.*, 2008) (unpublished results). From the 5,717 genes obtained for J6; 4,795 genes (~83.87%) matched to *D. fabryi* (CBS 789), 902 genes (~15.78%) matched *D. hansenii* (CBS 767) and 20 genes (~0.35%) matched other yeasts, such as *D. subglobosus*, *C. famata*, and proteins of unknown function. This demonstrates that at the protein level, *D. hansenii* J6 is more closely related to *D. fabryi* CBS 789 than *D. hansenii* CBS 767. Protein comparison against strain MTCC 234 could not be performed initially because protein sequences were not available in UniProt database. Thus, the MAKER pipeline was applied to the MTCC 234 contigs (public data set), using the same procedure as in the strain J6 (Berrocal *et al.*, 2016), to develop a protein profile that could be compared to J6 and other *Debaryomyes* species. This analysis showed 5,656 total genes for MTCC 234 compared to the 5,313 total genes reported by Kumar *et al.* for the same strain (2012). When compared to the UniProt Debaryomycetaceae family dataset, 4,801 genes (~84.8%) matched to CBS 789, 841 genes (~14.8%) matched to CBS 767 and 14 genes (~0.2%) matched to other yeasts. The protein matches done with MTCC 234 numerically concur with protein comparison analysis done for J6.

1.4 Pathogenicity

As mentioned earlier, *D. hansenii* was classified as an anamorph of the yeast *C. famata. D. hansenii* have been isolated from humans, specifically from interdigital mycotic lesions and from the throat of an angina and bone infection patients (Wong *et al.*, 1982). Previous investigations have demonstrated the phylogenetic similarity of *D. hansenii* to the yeasts *Candida guilliermondii* (ATCC 6260) and *Candida lusitaniae* (ATCC 42720), than other yeast species like *Candida albicans*, *Candida dubliniensis*, *Candida glabrata*,

Saccharomyces cerevisiae, Yarrowia lipolytica and *Kluyveromyces lactis* (Cai *et al.*, 1996, Fitzpatrick *et al.*, 2006, Woolfit *et al.*, 2007, Sherman *et al.*, 2009). Recently, a group of researchers investigated the difference in microbial communities on patients with cervical cancer, due to HPV (Human papillomavirus), from healthy individuals from Puerto Rico (Filipa Godoy, Personal communication, Paper In-Review). They found that woman with cervical cancer had samples with matches to the *Debaryomyces* genera. This demonstrates that *D. hansenii* is an opportunistic pathogen rather than a pathogen like *C. albicans*, which is the main organism in candidiasis infections (Odds, 1988) and is known to be an osmotolerant yeast species like *D. hansenii*. This highlights that the salt tolerance of the *Debaryomyces* genera may play an important role in the survivorship of these organisms in the human body, since NaCl concentrations may act as a barrier limiting the amount of organisms that can survive this environment. A study by Mattsson *et al.* showed that feral pigeons can be carriers of this yeast after they analyzed fresh droppings and cloaca samples from the bird *Columba livia* (1999), thus this yeast may disperse to many areas due to avian migration.

1.5 Biotechnology Applications

Due to its osmotolerance and wide range of growing temperatures, *D. hansenii* possess some advantages that other yeasts cannot offer. This yeast can be used in processes under high osmolality environments, limiting the use of chemicals to produce sterile batch processes, thus reducing the cost and time of production (Breuer and Harms, 2006). Also, *D. hansenii* has showed its importance in the agro-food industry like cheese making (Seiler and Busse, 1990), where its contribution is recognized because of the special flavors it can bring to the cheese (Ferreira and Viljoen, 2003). At the same time, *D. hansenii* can inhibit the growth of yeast species like *Clostridium butyricum and C. tyrobutyricum*, which are species not desired in cheese brines (Fatichenti *et al.*, 1983). In meats, *D. hansenii* produces ammonia and several volatile compounds, which modify the sensory properties of the meats (Flores *et al.*, 2004, Dura *et al.*, 2004, Andrade *et al.*, 2010). *D. hansenii* is also responsible for the production of the sweetener xylitol (Girio *et al.*, 1996, 2000) where some strains (*i.e.* NRRL Y-7426) have been improved and optimized (Parajo *et al.*, 1997). At the same time, some strains can produce arabinitol and

riboflavin in the growth phase on batch cultures (Anderson and Harris, 1963). Since *D. hansenii* exhibit anti-fungal activity, it could be used as a biocontrol agent against spoilage as demonstrated by Droby *et al.* (1999).

D. hansenii is also a lipid accumulating yeast, otherwise called an oleaginous organism. These organisms (*i.e.* oleaginous) are capable of accumulating lipids to concentrations up to 70% of their dry biomass (Ratledge and Tan, 1990). A group of researchers studied the lipid composition of *D. hansenii* and showed that around 67% consist of neutral lipids (Merdinger and Devine, 1965). These results are in concordance with experiments in our laboratory that have shown that some strains of *D. hansenii* can store neutral lipids when growing; making them good candidates for biodiesel production, by performing transesterification reactions to the lipids extracted from the yeast (unpublished results).

Some strains of this *D. hansenii* are flavinogenic, which produce high concentrations of riboflavin compounds in the presence of heavy metals (Gadd and Edwards 1986). Consequently, another application of *D. hansenii* could be as an organism for detecting heavy metals like cobalt. A study by Seda-Miró *et al.* demonstrated that strain J6, when exposed to toxic concentrations of cobalt, started producing riboflavin compounds in response to the heavy metal (2007). Their analysis consisted on monitoring the media color for the appearance of a yellow pigment indicating the contamination. Even though it is an interesting system, results could not be directly observed in the media until passed 72 hours after the exposure to cobalt (Seda-Miró *et al.*, 2007).

1.6 Previous studies for the selection of yeast for this research

Previous studies have demonstrated the great variability of different strains of *D. hansenii*. For instance, a study by Del Bove *et al.*, demonstrated that *D. hansenii* strains exhibit high molecular and metabolomic variability, and some strains that were before grouped together were differentiated by combination of different techniques (2009). Another study by Seda-Miró, evaluated the phenotypic variation of several *D. hansenii* strains isolated from different sources (Ph.D. Dissertation, 2012). They cultured 34 strains of *D. hansenii* under environmental stress with varying concentrations of Co (II)

(cobalt, oxidative stress) and NaCl (osmotic stress). The output of these experiments were classified as weakly tolerant, tolerant and highly tolerant. The differences they obtained in the phenotype characterization of these 34 strains, in response to cobalt and saline stress, prompted them to analyze the sequence variability of inducible genes.

For this purpose, they selected genes ENA1 and FET3 for further analysis. ENA1 is a gene induced (*i.e.* upregulated) by high-salt concentrations and encodes for an ATPase sodium pump (Ruiz and Ariño, 2007; Gorjan and Plemenitas, 2006); FET3 is a gene induced by the presence of cobalt and is associated with high-affinity iron transport systems (Stadler and Schweyen, 2002). After PCR amplification of these two genes in their 34 strains, a Restriction Fragment Length Polymorphism (RFLP) technique was applied using the enzymes EcoRI, BamHI, HindIII and Sau3A, obtaining six restriction profiles (Seda-Miró, Ph.D. Dissertation, 2012). These profiles were used for selecting the yeasts to be use for this study. We want to resolve the phylogeny of strains of *D. hansenii* taking one strain per observed pattern in the RFLP technique. Our hypothesis is that having representation of different groups (or clades), following the results of the RFLP technique, we would have a better understanding of the evolutionary history of the strains of *D. hansenii*. We will use other currently available genomes of *D. hansenii* and *D. fabryi* to test the possibility of wrong taxonomic assignment of this species.

2. MATERIALS AND METHODS

2.1 Culturing and extractions

For this research, we selected *D. hansenii* strains: NRRL Y-1449, NRRL Y-155, NRRL Y-398, NRRL Y-17914, NRRL Y-j26 and NRRL Y-7426 (Table 1). These strains resulted in different patterns when a RFLP technique was applied to them (see section 1.6). Each strain was grown in 250 mL of YPD medium (1% yeast extract, 2% peptone, 2% dextrose) and incubated in an orbital shaker (150 rpm) at 25°C. Yeast cells were collected by centrifugation followed by the storage of cell pellets in -80ºC. DNA extractions were done using the method of Cryer *et al*. (1975) followed by gel electrophoresis to corroborate the extraction was successful.

Table 1. Strains use in this study.

Strain	Source of isolation	Cuture collection	Genome creation and annotation
Y-1449	Throat of angina patient, France	NCAUR	This study
YB-155	Coconut fruit salad	NCAUR	This study
Y-17914	Interdigital mycotic lesion, Germany	NCAUR	This study
Y-7426	Carlsberg Research Lab, Denmark	Dr. C. Kurtzman, National Center for Agriculture Utilization Research (NCAUR), US Department of Agriculture, Peoria, Ill.	This study
J26	Swedish estuary	University of Gothenburg, Sweden	This study
YB-398	High moisture corn	NCAUR	This study
CBS 767	Carlsberg Research Lab, Denmark	Dr. Bernard Dujon (Institut Pasteur and University Pierre & Marie Curie-Paris)	Dujon *et al*., 2004
MTCC 234	New Zealand soil	MTCC	Kumar *et al*., 2012; This study
CBS 789	Human Interdigital Mycotic Lesion	CBS	Tafer *et al*., 2016
J6	Swedish estuary	University of Gothenburg, Sweden	Berrocal *et al*., 2016
C.guilliermondi ATCC 6260	Sputum from patient with bronchomycosis	ATCC	Butler *et al*., 2009

2.2 Sequencing

DNA samples were sent to the Hudson Alpha Genome Sequencing Center in Alabama for library preparation and sequencing. The six libraries were sequenced using the Illumina® HiSeq2000 platform with paired-end reads of 100bp of length (2x100bp).

2.3 Quality check, mapping and *de novo* assemblies

Reads obtained from the sequencing center (Table 2) were checked for quality and possible adaptor presence using FastQC (Andrews, 2010). Reads were subsequently trimmed using a custom Cutadapt script for elimination of sequencing adaptor, reads with less that 15 in quality scores (phred scores) and the first 5 bases of the sequences (Martin, M. 2011). Using Bowtie2 (Langmead, B., and Salzberg, S. L., 2012), reads were mapped to other available genomes of *D. hansenii* and *D. fabryi* (*i.e.* CBS 767, MTCC 234, CBS 789T, J6) (Dujon *et al*., 2004, Kumar *et al*.,2012, Tafer *et al*., 2016, Berrocal *et al*., 2016; respectively). *De novo* assemblies of genomes were also performed using the program ABySS (Simpson *et al.*, 2009), with different K-mer values from 31 to 81 for each strain using custom bash scripts. For this step, we used the Greenfield system at the Pittsburgh Supercomputing Center (PSC), under the project Extreme Science and Engineering Discovery Environment (XSEDE)(Towns *et al.*, 2014). All contigs less than 200bp were removed from all the assemblies. Quality analysis of each filtered assembly was made by QUAST (Gueverich *et al*., 2013). We used the N50 (*i.e*. base pairs were 50% of the total length of the assembly can be found) values to measure the quality of genome assembly (Yandell and Ence 2012).

2.4 Gene prediction and annotation

Gene prediction and annotation of each strain were performed using the MAKER pipeline (Cantarel *et al.*, 2008), with both transcriptomics and genomics data generated from our lab and/or publicly available. All the proteins available on the UniProt database from the phylum Ascomycota were used for the annotation process. This process of gene prediction and annotation is crucial for the selection of genes for the phylogenetic analysis (section 2.5).

2.5 Selection of genes for phylogeny generation

The six *D. hansenii* genomes created in this study (see sections above) and five public available genomes (J6, CBS 767, CBS 789(*fabryi*), MTCC 234, *Candida guilliermondi* ATCC 6260) were used for the phylogenetic analysis. For selecting the genes, we developed a custom bash script (see appendix), taking advantage of the rigorous annotation structure by the MAKER pipeline.

First, the program searched for each protein of *D. hansenii* CBS 767 (type strain) in each genome annotation created with an exception to the strain NRRL Y-17914. The reason for skipping this strain is because it is more like *D. fabryi* CBS 789 than CBS 767 after we performed the annotation (see Results), and it did not match the nomenclature used for the annotation procedure with strain CBS 767. The process to obtain the genes for NRRL Y-17914 is discussed below. Then, each protein that was shared by all the strains (*i.e.* name) was saved in a new file. This file was used to extract each sequence from the annotation using custom python scripts (see appendix). Any sequence with a duplicate was stored in a separate file for further analysis.

Basic Local Alignment Search Tool (Blast) protein searches are performed using the sequences of all shared genes, from the previous step (using CBS 767 sequences), as a database. The genomes of *C. guilliermondi* ATCC 6260 and *D. fabryi* CBS 789 were used as queries for the search. The result of the Blast search of CBS 789 was used to get the shared genes of the strains J6, MTCC 234 and NRRL Y-17914, because all these strains matched to CBS 789 during the annotation process. The Blast search done to *C. guilliermondii* (outgroup organism) was used to get the shared genes between *C. guilliermondi* and all the other *Debaryomyces* strains. After the script is done, all the shared genes between all the genomes in this study (*i.e.* 11) were concatenated together in separate files, each file with the name of the corresponding gene. These files are then used for phylogenetic analysis (see section 2.6).

2.6 Phylogenetic analysis

Two methods were applied for the phylogenetic analysis, concatenation and coalescence. In the concatenation methodology, alignments created of each gene are concatenated together to then form a super alignment. This file is then analyzed and a tree that resolves the phylogeny is generated. The coalescence method first generates a tree for each of the gene in the study. Then a final tree is created using this information (*i.e.* individual trees) as an input, thus creating what is called a super tree. The program RAxML (Stamatakis, 2014) was used for the Maximum Likelihood analysis and MrBayes (Ronquist *et al*., 2012) for the Bayesian analysis. In both procedures, we used MAFFT aligning program (version 7.2) (Katoh and Standley, 2013) with iterative refinement methods using WSP and consistency scores (*i.e.* G-INS-i). This method is generally slow but its accuracy is better than any other method applied in MAFFT (Katoh and Standley, 2013). It works best with less than 200 sequences with global homology, as in the present study. Each gene will have a total of 11 sequences, one for each yeast species, thus it is within the requirements of the MAFFT program (*i.e.* G-INS-i).

2.6.1 Concatenation

For the concatenation procedure, we aligned the sequences in each gene file generated with the custom bash script we developed (see section 2.4) using the program MAFFT (G-INS-i). Each alignment was concatenated in a new file, using another custom bash script (see appendix), to form a matrix of all the alignments together for each strain of the study. This matrix file was then aligned again with MAFFT to create the final super alignment, but with the auto option (*i.e.* selection of the best method to align sequences depending on dataset size) instead of G-INS-i. Using a perl script developed by Dr. Shannon Hedtke (La Trobe University), we converted the super alignment to a nexus format to be used in MrBayes. We then used MrBayes with the GTR + GAMMA (Tavaré 1986, Yang 1994, 1996) model, to account for amino acid substitution and rate heterogeneity among the sites. We ran the program with 10,000,000 generations, a sample frequency of 1000 and the save tree option activated. All other parameters were left in default. We also used RAxML with the rapid bootstrapping method (-f a) and 1000

bootstraps with the GTR + GAMMA model to compare both results (see section 3). GTR + GAMMA model was used in both methodologies because it includes all possible amino acid substitution rate, thus making an exhaustive analysis of the sequences. All trees were visualized using FigTree (cite).

2.6.2 Coalescence

For the coalescence analysis, we used the program ASTRAL-II (Mirarab and Warnow, 2015). This program takes as input unrooted trees to then perform the analysis. Each gene from this study was aligned using MAFFT. We then incorporated the java program readseq (version 2.1.30) developed by Don Gilbert (Indiana University) to convert each individual gene alignment into a Phylip|Phylip4 format. These alignments in phylip format were used in the program RAxML to create individual trees that are going to be used in the program ASTRAL-II. We ran RAxML using the rapid bootstrapping method (-f a) with 1000 bootstraps and the GTR + GAMMA model. All individual gene trees and its bootstraps replicates were used to resolve the phylogeny of the *Debaryomyces* species complex. All the parameters in ASTRAL-II were left in default with 100 replicates and full annotation (-t 2). Final trees were visualized using FigTree.

3. RESULTS

3.1 Sequencing and filtering

Reads obtained from the sequencing center are presented in Table 2. An average of 16,166,904 reads (max=17,597,330 (strain Y-1449); min=14,579,466 (strain Y-17914)) per strain sequenced were obtained. Of these sequenced reads, only an average of 13,694,666 met the criteria of the cutadapt scripts (see section 2.3). We retained more than 80% of the sequenced reads for the creation of the genomes (Table 2).

Table 2. Reads obtained from the sequencing center and after filtering

Strain	Reads after sequencing	Reads after filtering	Percent of reads retained
Y-1449	17,597,330	14,912,412	84.7
YB-155	17,063,399	14,772,937	86.6
Y-17914	14,579,466	12,600,809	86.4
Y-7426	15,495,416	13,327,969	86
J26	16,619,997	13,832,261	83.2
YB-398	15,645,816	12,721,607	81.3
Average	16,166,904	13,694,666	84.7

3.2 Mapping

Reads that successfully passed quality check were mapped to existing genomes of other *Debaryomyces* yeasts. Percentage of mapping and the yeasts used as reference (*i.e.* best mapping result obtained) are presented in Table 3. Five of the six yeasts created in this study had better mapping percentage when mapped to strain CBS 767 (species *hansenii*). However, the remaining strain, Y-17914, was the only strain that showed higher similarity to strain CBS 789, which is a member of the species *fabryi* (Table 3). Of these mapping results, four strains mapped to strain CBS 767 at around 50%. The other two strains mapped at around 74%. These two strains were NRRL-Y 7426 with CBS 767 (75.45%) and NRRL-Y17914 with CBS 789 (73.54%).

Table 3. Mapping results of the strains created in this study and their reference strain.

Strain	Percentage of mapping	Reference strain
Y-1449	51.87	CBS 767
YB-155	51.35	CBS 767
Y-17914	73.54	CBS 789
Y-7426	75.45	CBS 767
J26	48.59	CBS 767
YB-398	50.71	CBS 767

3.3 *De novo* assemblies

Since the mapping results for some species were not high enough and to investigate further our sequences, we decided to perform *de novo* assemblies for each strain in this study. Table four summarizes genome size, number of scaffolds, N50 values, GC content, and putative unique genes for each *de novo* assembled strain. These results were obtained using the program QUAST. This program was used in each *de novo* assembly for each strain (*i.e.* 25 assemblies in total per strain)(see section 2.3). The quality of genome assembly was evaluated using N50 values as explained in the method section. From all the assemblies performed to each strain, we selected the assembly with higher N50 values. The strain with a minimum value of N50 from all the assembled strains was J26 with 101,137 bp; and the strain with a maximum value of N50 was for strain Y-17914 with 233,948 bp. The average N50 for all the assemblies was 163,008 bp (Table 4).

Table 4. Results from the genome assemblies with the MAKER program.

Strain	Genome size	Number of scaffolds after filtering (>200bp)	N50	GC content	Putative genes (unique)
Y-1449	12.02 Mb	374	170,318 bp	36.29%	5664
YB-155	12.34 Mb	346	176,737 bp	36.17%	5802
Y-17914	11.91Mb	263	233,948 bp	35.65%	5664
Y-7426	12.14 Mb	353	183,597 bp	36.30%	5786
J26	12.19 Mb	496	101,137 bp	36.30%	5748
YB-398	12.12 Mb	405	112,309 bp	36.32%	5718

3.4 Gene prediction and annotation

We applied the MAKER pipeline to develop gene prediction and annotation. After the analysis was concluded we obtained an average of 5904 predicted genes in the *D. hansenii* strains. Of these predicted genes, most were annotated using the strain CBS 767 as reference in the protein similarity search with an exception of strain Y-17914 (Table 5). These annotated genes were then used as input for the custom bash program to select the shared genes between the strains. Predicted transfer RNAs (tRNAs) are also presented in Table five.

Table 5. Predicted genes for each strain and the reference used in the genome annotation.

Strain	Predicted genes	Predicted tRNA	CBS 767	CBS 789	Other yeasts
Y-1449	5851	187	5549	257	45
YB-155	5932	219	5374	508	50
Y-17914	5844	224	491	5324	29
Y-7426	5956	230	5781	143	32
J26	5927	161	5610	273	44
YB-398	5915	197	5596	260	59

3.5 Genes for phylogeny generation

Table 6 summarizes the genes shared between each strain of this study and the type strain CBS 767. Of the more than 5900 genes (in average) that were predicted and annotated in each strain we assembled, 4972 genes were shared between all the species on Table 6 and CBS 767. From those genes, only 4797 were represented one time in the file (single copy genes), thus they were selected for further analysis. The remaining 175 genes that did not pass this filtering step may be multiple copy genes from some species, as a consequence these genes were discarded from our analysis.

Table 6. Shared genes between the strains and CBS 767.

Strain	Shared genes with CBS 767
Y-1449	5386
YB-155	5182
Y-7426	5636
J26	5396
YB-398	5414
Shared between all the above strains	4972

After the Blast search, the genes that were shared between CBS 767, CBS 789 (*D. fabryi*) and ATCC 6260 (*C. guilliermondii*) were 4307. From these genes, the script searched for how many of these genes were shared between CBS 789, MTCC 234 and Y-17914, resulting in 3649, 3649 and 4014, respectively. These genes were then extracted and the script searched for the shared genes between all the strains (*i.e.* created in this study and publicly available, n=11) resulting in a total of 3578 shared genes after using *C. guilliermondii* (ATCC 6260) as the outgroup species. We also repeat this process with *C. albicans* (ATCC MYA-2876) and *C. luisitinae* (ATCC 42720) resulting in 3249 and 3513 genes, respectively. The 3578 genes obtained using *C. guilliermondii* were then used for the phylogeny analysis.

3.6 Phylogenetic analysis

3.6.1 Concatenation

The 3578 genes obtained from the custom bash program were aligned and concatenated for further analysis. A RAxML tree presented a reconstruction of the phylogeny with support values for each branch of one hundred (Figure 1). The genetic change (or substitution rate) obtained was 0.20. The topology of the tree followed the pattern observed in the mapping and annotation procedure where strain CBS 767 matched with NRRL-Y 7426, and strain CBS 789 matched with NRRL-Y 17914. Similar results were obtained for the concatenation tree using the program MrBayes (Figure 2). The topology of the tree was the same as the tree obtained by RAxML, and each branch has support

values of one hundred. However, the genetic change observed was lower, 0.07, when compared to the change obtained by RAxML. The grouping of strains NRRL-Y 7426 with CBS 767, and strains NRRL-Y17914 with CBS 789 was also observed. Both trees grouped strain CBS 789 (*fabryi* species) together with *hansenii* species. As expected, *C. guillermondii* (*i.e.* outgroup) is the organism with more genetic difference as observed by the length of the branch (Figure 1 and 2).

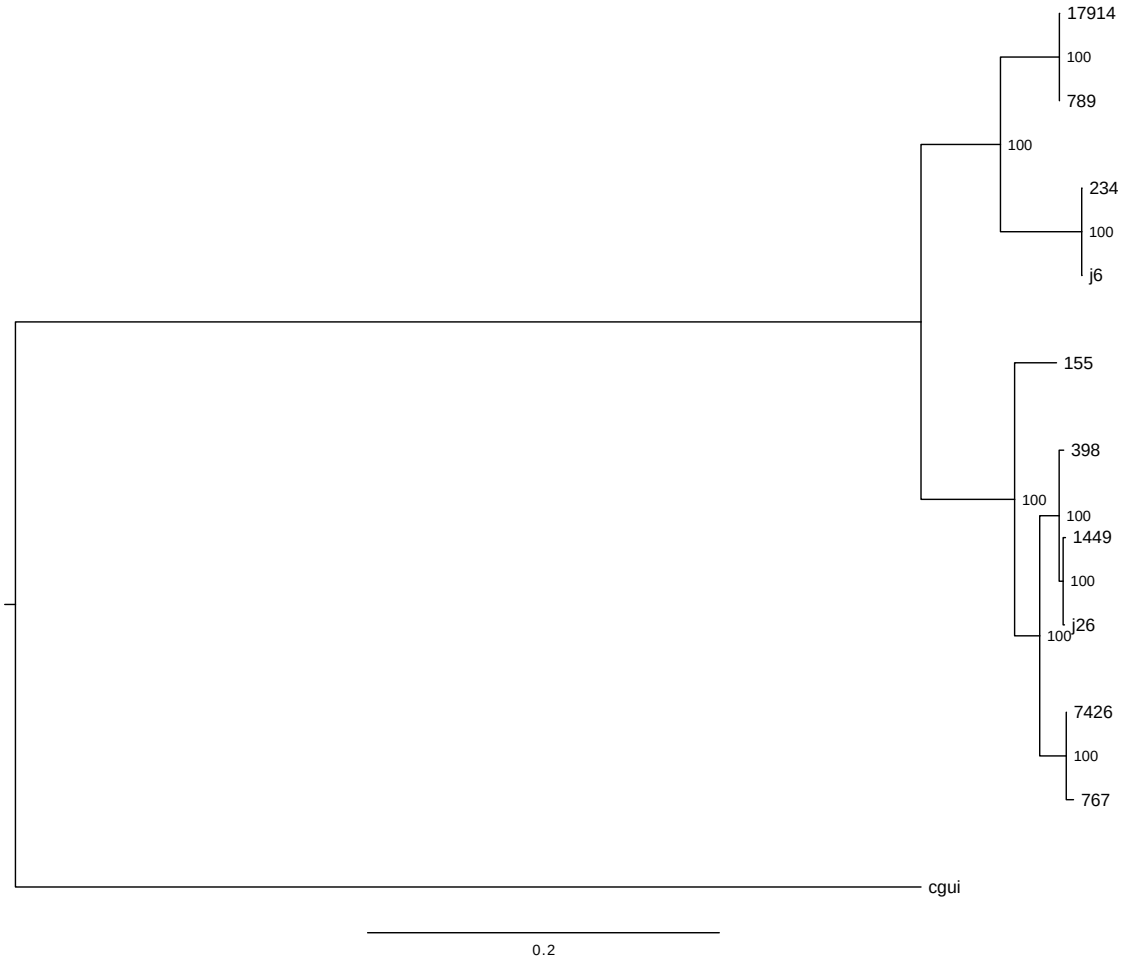

Figure 1. RAxML concatenation tree of the shared genes between all the species in the study (*i.e* 11 species, 3578 genes)

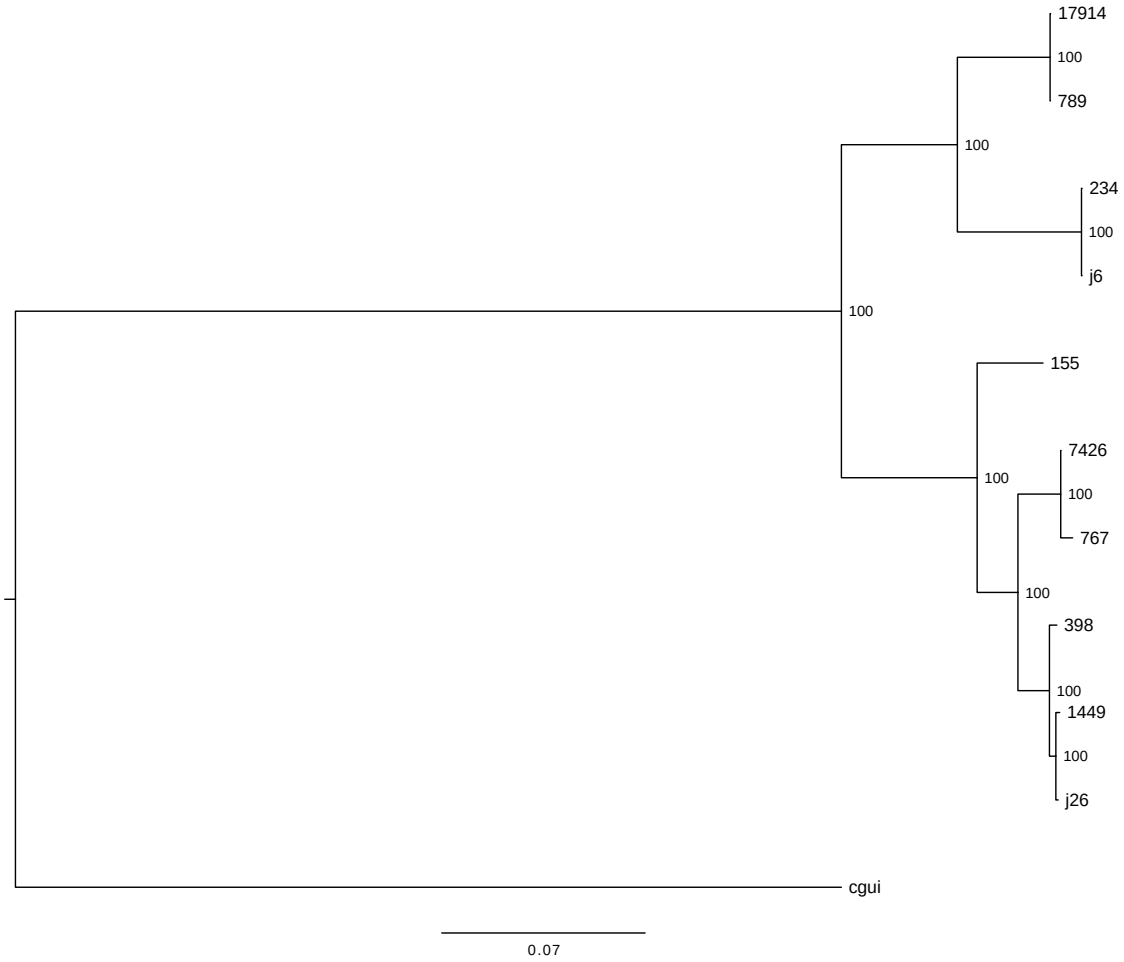

Figure 2. MrBayes concatenation tree of the shared genes between all the species in the study (*i.e* 11 species, 3578 genes)

3.6.2 Coalescence

The tree obtained in the coalescence analysis had similar topology with the trees of the concatenation analysis (Figure 3). The strain-grouping pattern observed in the concatenation trees was also observed (strains NRRL-Y 7426 with CBS 767, and strains NRRL-Y17914 with CBS 789). Support values for each branch were also one hundred. The genetic change observed was 2.0 but expressed as coalescence units. As expected, *C. guillermondii* (*i.e.* outgroup) is also the organism with more genetic difference, a similar result in the concatenation trees.

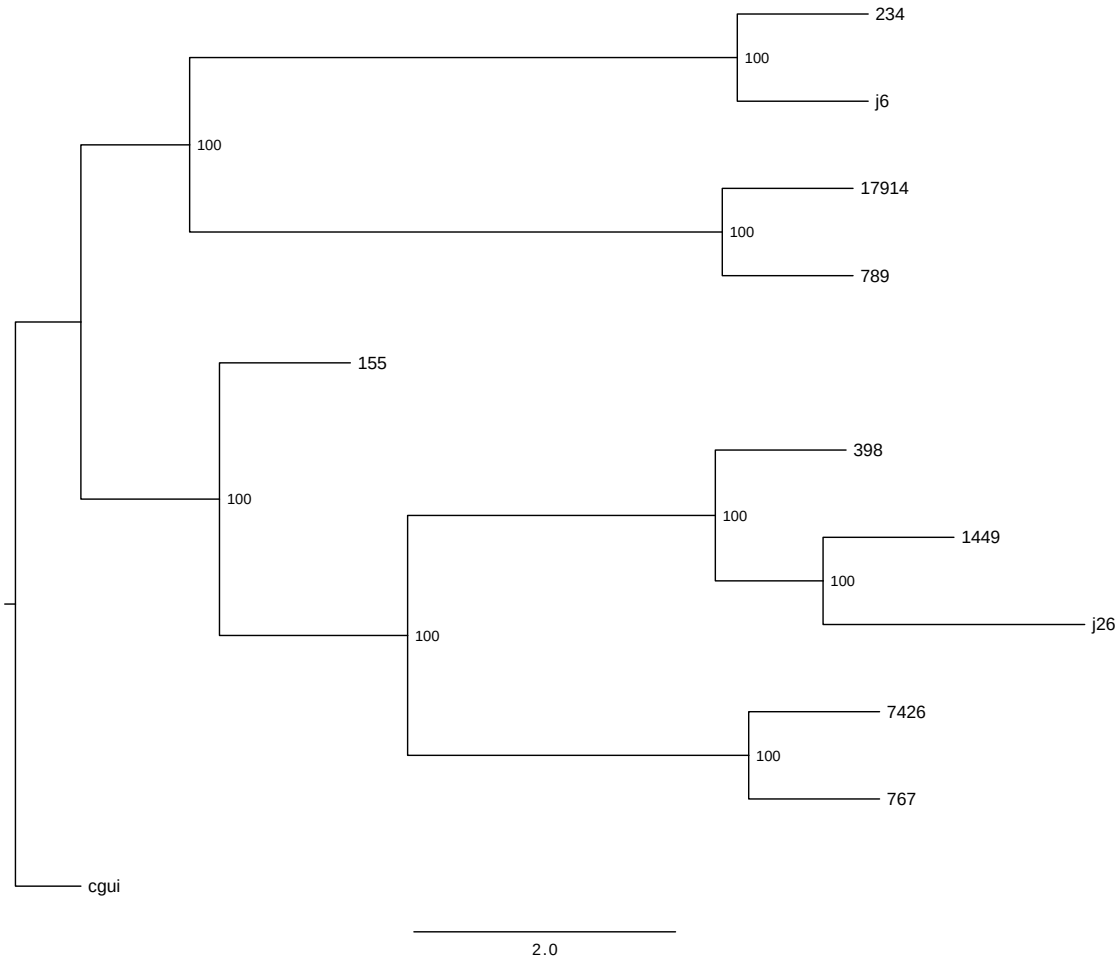

Figure 3. Coalescence tree generated by the program ASTRAL-II for the shared genes between all the species in the study (*i.e* 11 species, 3578 genes).

4. Discussion

The purpose of this research was to apply NGS technology for the phylogenetic study of the *Debaryomyces* spp. complex. After sequencing, a series of steps were done to ensure a good assembly of each sequenced strain. Reads were trimmed and possible contamination was removed. Mapping to existing genomes showed that most strains mapped to strain CBS 767 (*hansenii* species) (Table 3). The only strain that did not map with CBS 767 was strain NRRL-Y 17914, which had better mapping to the strain CBS 789 (*fabryi* species) (~74% of mapping). These results were expected since all these strains that mapped to CBS 767 were classified as a *D. hansenii* var. *hansenii*, and strain NRRL-Y 17914 was classified as *D. hansenii* var. *fabryi* (Seda-Miro *et al.,* 2012). This indicates that this species maybe wrongly classified as *D. hansenii* and not as *D. fabryi*. This exact pattern was also observed in the annotation procedure where strain CBS 789 showed higher similarity to strain NRRL-Y 17914. From the inputted proteins use for the annotation procedure, proteins from strain CBS 767 were used as reference to annotate five of the six species from the study (Table 5). The remaining strains annotated using strain CBS 789 as a reference, presented the same pattern obtained from the mapping procedure (Table 3). These two results are important when analyzing and reconstructing the evolutionary history of these species, because it gives insight into the relationship of the species before the phylogeny is constructed and can be use two asses the accuracy of the tree.

Given these differences between the strains and that some mapping percentages were relatively low (*i.e.* ~50%), we performed *de novo* assemblies of each strain. The results of this step are like previous sequencing projects (Dujon *et al.,* 2004). In this study, they sequenced the first yeast genome of *D. hansenii* (CBS 767). This strain has a genome of 12.2 million bases (Mb) with a GC content of 36.3%, and encodes for 6,906 genes. The assembled genomes we developed in this study have a size range between 11.91 to 12.34 Mb and have GC content between 35.65 to 36.32%, which are similar to the strain CBS 767. In the gene predictions and annotation, most genomes had an average of 5900 genes.

This result differed when compared to the results obtained for strain CBS 767 (Dujon *et al.*, 2004). Although there is a difference of ~1,000 genes, we believe our assembly and genes prediction are correct. Our sequence depth was over 150%, which is necessary to ensure good downstream processing. Also, in the gene prediction and annotation procedure, we employed as much information as possible. We used transcriptome data from *D. hansenii* (Guma-Cintron *et al.*, 2015), genomes from related species (Berrocal *et al.*, 2016), all the proteins from the Phylum Ascomycota on UniProt, together with different gene prediction tools. Some of these tools were already tailor-made for being used with *D. hansenii* species (*i.e.* Augustus). Also, *D. hansenii* strains have been showed to be very heterogeneous organisms, so expecting this species to have exact number genes is not realistic. We are not suggesting our genome is complete and maybe some shallow sequencing would be needed to ensure a complete genome, however given the overall results of genome size and number of genes, we estimate that we have over 90% of the genome.

All the genes obtained from the annotation were used for the selection of the best candidate genes for building the phylogeny. As demonstrated on Table six, five of the six species from this study had the majority of the genes annotated using strain CBS 767 as a reference. Taking the average gene number of each strain of the study (*i.e.* 5900), the shared genes between these five species is around 84% (*i.e.* 4972 genes) of the total genes from the reference strain. After searching for the corresponding genes of CBS 789, *C. guilliermondii* (ATCC 6260), and all the remaining strains from the study, we obtained 3578 shared genes between all the strains from this study (*i.e.* 11). This number of genes is comparable to other studies that worked on the phylogeny of yeast species (Shen *et al.*, 2016, Fitzpatrick *et al.*, 2006). However, to be as rigorous as possible in the selection of the candidate genes, we only took the genes that were present in all the species and did not included genes that may present in all the species except for one or two (*i.e.* 100% of gene occupancy). This may had limited the number of genes for the phylogeny generation, but since we may not have complete genomes we do not want to bias or limit our analysis.

After alignment with MAFFT, all the genes were concatenated together before analysis using Maximum Likelihood and Bayesian statistics for the concatenation procedure. From the Maximum Likelihood tree (Figure 1) we observed the grouping pattern obtained from previous bioinformatics steps (*e.g.* mapping). Species CBS 767 and NRRL-Y 7426 grouped together, and strain CBS 789 grouped with strain NRRL-Y 17914. This grouping of yeast species was also observed for the Bayesian tree (Figure 2). Both trees topologies were identical, including the outgroup species assignment and support values of 100% for all branches. In the case of the coalescence analysis, our tree presented the same topology as the concatenation trees (Figure 3), and support values of 100%.

Since all the trees had the same topology, we decided to create a cladogram for easy analysis of the evolutionary relationship between the species of the study (Figure 4). A peculiar arrangement of the species in all trees was observed and prompted us to further investigate these species. Our tree is divided into two main branches when excluding the outgroup species (see color boxes on Figure 4). Group A (red box) consisted of species currently classified as *Debaryomyces hansenii*. Group B (blue box) consisted of a mixture of species classified as *Debaryomyces hansenii* or *Debaryomyces fabryi*. Taxonomically speaking, some species from group B may have been wrongly assigned. For instance, strains CBS 789 and Y-17914 are classified as *D. fabryi* and *D. hansenii*, respectively (Figure 4). We cannot say which of these species is wrongly assigned with confidence, since this problem of taxonomic identification has been continuously occurring. However, we think strain Y-17914 maybe wrongly assigned because in some studies this species has been classified as *D. hansenii* var. *fabryi* (Seda-Miro *et al.*, 2012).

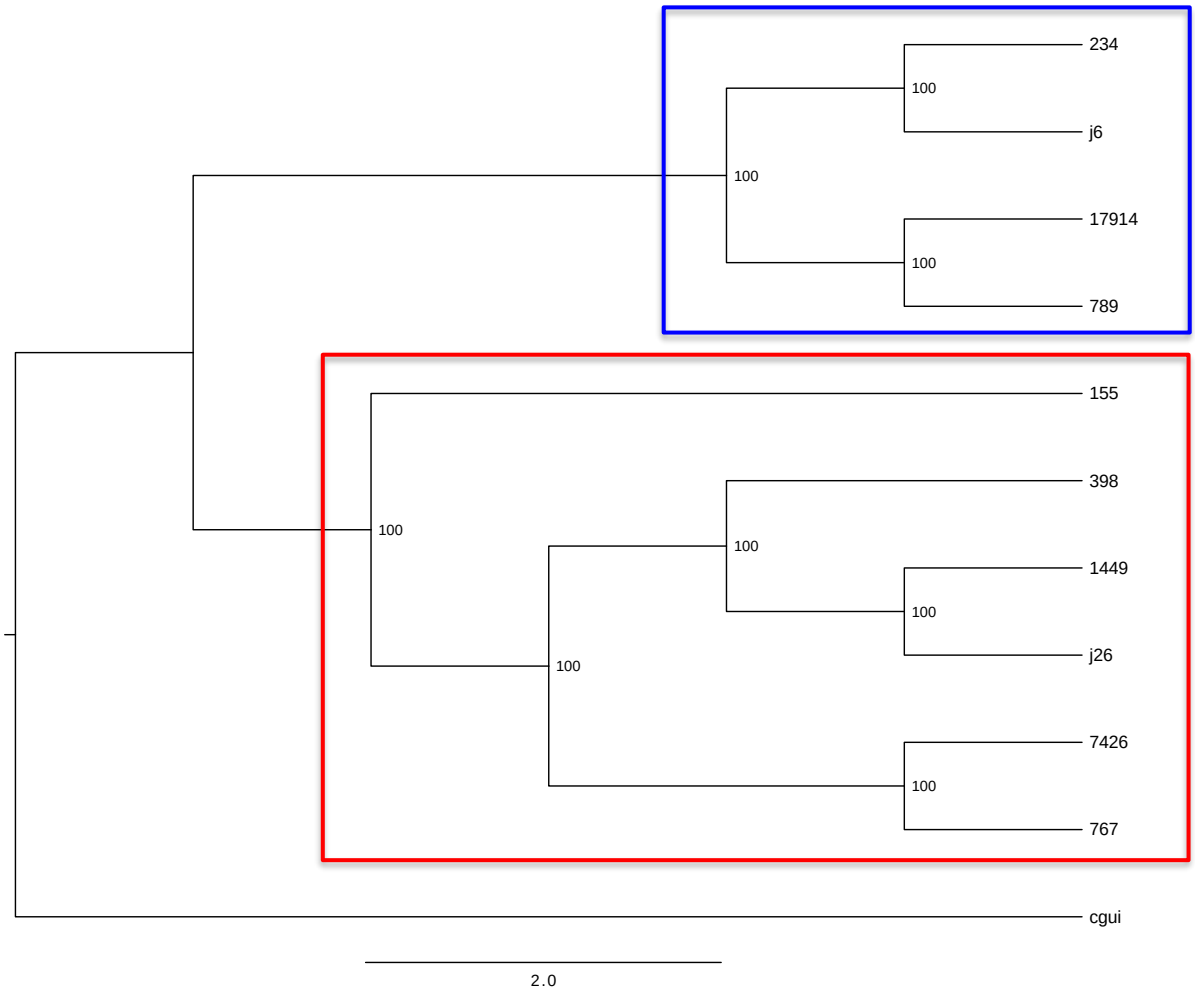

Figure 4. Cladogram generated using the information of all the previous trees on this study. Red box consisted of species currently classified as *Debaryomyces hansenii*. Blue box consisted of a mixture of species classified as *Debaryomyces hansenii* or *Debaryomyces fabryi*.

The division of the yeast species by two main branches also points out an interesting trait used for classification (Figure 4). The species of group B, all possess the capability of producing riboflavin compounds when exposed to heavy metals (*i.e.* flavinogenic species). This result may suggest a speciation event that took place during an early evolutionary period, which divided the ancestor species that are flavinogenic from ancestor species that are non-flavinogenic. Flavinogenic capability was suggested before as a method to divide species by previous researchers (Nguyen *et al.*, 2009). Although

more species and evidence may be needed to support this speciation event, we suggest that this capability may be used to divide species instead of relying only on differentiation techniques like maximum growth temperature (Groenewald *et al.*, 2008; Nguyen *et al.*, 2009). Newly found species should be tested for riboflavin production, followed by genome sequencing for the generation of phylogenies. In this way we can have a clear understanding of the evolutionary history of this newly found species and we may test is this capability should be used as a distinctive trait for differentiation.

This speciation event does not give more insight on how species of *D. hansenii* and *D. fabryi* are grouped together. However, it can be hypothesized, that this speciation event (*i.e.* flavinogenic vs. non-flavinogenic) took place before other traits evolved, resulting in the species differentiation of *D. hansenii* and *D. fabryi* species. For this conclusion, we need to sequence more genomes of *D. hansenii* and *D. fabryi*, with some species being flavinogenic. We need to ensure that our conclusions are based on a diverse group of species. Currently, we have six species on group A (*i.e.* red box, Figure 4) that are non-flavinogenic versus four species that are flavinogenic on group B. We suggest sequencing of at least four more flavinogenic yeast genomes, in which three should be from the species *fabryi* and one from the species *hansenii*. By doing this we can test if the speciation event of flavinogenic capabilities evolved before any strain differentiation (*i.e. D. hansenii* and *D. fabryi*). Another possibility that cannot be discarded is the wrong taxonomic assignment of the species. Like in the case of strain CBS 789, which is assigned as a *fabryi* species instead of *hansenii* species. However, as mentioned earlier more sequencing needs to be performed in order to have well supported analysis.

In conclusion, this is the first genomic research that presents a possible speciation event of the flavinogenic capability, and how it may be relevant in discerning taxonomic classification of yeast species from the genera *Debaryomyces*. This work demonstrates that the use of genomic information, for the construction of phylogenies, can be used to taxonomically assign yeast species from closely related groups. With the ease and feasibility of genome sequencing projects more insight and valuable information can be obtained to resolve species complexes such as in the *Debaryomyces* spp. complex.

References

Adler, L. (1986, October). Physiological and biochemical characteristics of the yeast Debaryomyces hansenii in relation to salinity. In The Biology of Marine Fungi. 4th International Marine Mycology Symposium. Portsmouth, England. Cambridge University Press, Cambridge, England (pp. 81-90).

Aggarwal, M., & Mondal, A. K. (2006). Role of N-terminal hydrophobic region in modulating the subcellular localization and enzyme activity of the bisphosphate nucleotidase from Debaryomyces hansenii. *Eukaryotic cell*, *5*(2), 262-271.

Aggarwal, M., Bansal, P. K., & Mondal, A. K. (2005). Molecular cloning and biochemical characterization of a 3′(2′), 5′-bisphosphate nucleotidase from Debaryomyces hansenii. *Yeast*, *22*(6), 457-470.

Almagro, A., Prista, C., Benito, B., Loureiro-Dias, M. C., & Ramos, J. (2001). Cloning and expression of two genes coding for sodium pumps in the salt-tolerant yeast Debaryomyces hansenii. *Journal of bacteriology*, *183*(10), 3251-3255.

Almagro, A., Prista, C., Castro, S., Quintas, C., Madeira-Lopes, A., Ramos, J., & Loureiro-Dias, M. C. (2000). Effects of salts on Debaryomyces hansenii and Saccharomyces cerevisiae under stress conditions. *International journal of food microbiology*, *56*(2), 191-197.

Anderson, F. B., & Harris, G. (1963). The production of riboflavin and D-arabitol by Debaryomyces subglobosus. *Microbiology*, *33*(1), 137-146.

Andrade, M. J., Córdoba, J. J., Casado, E. M., Córdoba, M. G., & Rodríguez, M. (2010). Effect of selected strains of Debaryomyces hansenii on the volatile compound production of dry fermented sausage "salchichón". *Meat Science*, *85*(2), 256-264.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Berrocal, C. A., Rivera-Vicens, R. E., & Nadathur, G. S. (2016). Draft genome sequence of the heavy-metal-tolerant marine yeast Debaryomyces hansenii J6. *Genome Announcements*, *4*(5), e00983-16.

Breuer, U., & Harms, H. (2006). Debaryomyces hansenii—an extremophilic yeast with biotechnological potential. Yeast, 23(6), 415-437.

Butinar, L., Santos, S., Spencer-Martins, I., Oren, A., & Gunde-Cimerman, N. (2005). Yeast diversity in hypersaline habitats. *FEMS Microbiology Letters*, *244*(2), 229-234.

Butinar, L., Strmole, T., Gunde-Cimerman, N. (2011). Relative incidence of ascomycetous yeasts in arctic coastal environments. *Microbial ecology*, *61*(4), 832-843.

Butler G, Rasmussen MD, Lin MF, Santos MA, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, Agrafioti I, Arnaud MB, Bates S, Brown AJ, Brunke S, Costanzo MC, Fitzpatrick DA, de Groot PW, Harris D, Hoyer LL, Hube B, Klis FM, Kodira C, Lennard N, Logue ME, Martin R, Neiman AM, Nikolaou E, Quail MA, Quinn J, Santos MC, Schmitzberger FF, Sherlock G, Shah P, Silverstein KA, Skrzypek MS, Soll D, Staggs R, Stansfield I, Stumpf MP, Sudbery PE, Srikantha T, Zeng Q, Berman J, Berriman M, Heitman J, Gow NA, Lorenz MC, Birren BW, Kellis M*, Cuomo CA*. (2009). Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature, 459(7247), 657.

Cai, J., Roberts, I. N., & COLLINS, M. D. (1996). Phylogenetic relationships among members of the ascomycetous yeast genera Brettanomyces, Debaryomyces, Dekkera, and Kluyveromyces deduced by small-subunit rRNA gene sequences. *International Journal of Systematic and Evolutionary Microbiology*, *46*(2), 542-549.

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., ... & Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome research, 18(1), 188-196.

Ciccarelli, F.D., T. Doerks, C. von Mering, C.J. Creevey, B. Snel and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311, 1283–1287.

Cong, Y. S., Yarrow, D., Li, Y. Y., & Fukuhara, H. (1994). Linear DNA plasmids from Pichia etchellsii, Debaryomyces hansenii and Wingea robertsiae. *Microbiology*, *140*(6), 1327-1335.

Corredor, M., Davila, A. M., Casarégola, S., & Gaillardin, C. (2003). Chromosomal polymorphism in the yeast species Debaryomyces hansenii. Antonie van Leeuwenhoek, 83(3), 215-222.

Corredor, M., Davila, A. M., Gaillardin, C., & Casaregola, S. (2000). DNA probes specific for the yeast species Debaryomyces hansenii: useful tools for rapid identification. FEMS microbiology letters, 193(1), 171-177.

Cryer DR, Eccleshall R, Marmur J. 1975. Isolation of yeast DNA. Methods Cell Biol 12:39 – 44. http://dx.doi.org/10.1016/S0091-679X(08)60950-4.
Davenport, R. R. (1980). Cold-tolerant yeasts and yeast-like organisms. *Biology and activities of yeasts. Academic Press, London*, 215-230.

de Silóniz, M. I., Valderrama, M. J., & Peinado, J. M. (2000). A chromogenic medium for the detection of yeasts with β-galactosidase and β-glucosidase activities from intermediate moisture foods. *Journal of food protection*, *63*(5), 651-654.

Del Bove, M., Lattanzi, M., Rellini, P., Pelliccia, C., Fatichenti, F., & Cardinali, G. (2009). Comparison of molecular and metabolomic methods as characterization tools of Debaryomyces hansenii cheese isolates. *Food microbiology*, *26*(5), 453-459.

Desnos-Ollivier, M., Ragon, M., Robert, V., Raoux, D., Gantier, J. C., & Dromer, F. (2008). Debaryomyces hansenii (Candida famata), a rare human fungal pathogen often misidentified as Pichia guilliermondii (Candida guilliermondii). Journal of clinical microbiology, 46(10), 3237-3242.

Droby, S., Lischinski, S., Cohen, L., Weiss, B., Daus, A., Chand-Goyal, T., ... & Manulis, S. (1999). Characterization of an epiphytic yeast population of grapefruit capable of suppression of green mold decay caused by Penicillium digitatum. *Biological control*, *16*(1), 27-34.

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuvéglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E, Bleykasten C, Boisramé A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L, Fabre E, Fairhead C, Ferry- Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud JM, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard GF, Straub ML, Suleau A, Swennen D, Tekaia F, Wésolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, Souciet JL. (2004). Genome evolution in yeasts. *Nature*, *430*(6995), 35.

Durá, M. A., Flores, M., & Toldrá, F. (2004). Effect of growth phase and dry-cured sausage processing conditions on Debaryomyces spp. generation of volatile compounds from branched-chain amino acids. *Food chemistry*, *86*(3), 391-399.

Eliskases-Lechner, F., & Ginzinger, W. (1995). The yeast flora of surface-ripened cheeses. *Milchwissenschaft (Germany)* 50, 458–462.

Fatichenti F, Bergere JL, Deiana P, Farris GA. (1983). Antagonistic activity of *Debaryomyces hansenii* towards *Clostridium tyrobutyricum* and *C. butyricum*. *J Dairy Res* **50**: 449 – 457.

Ferrando, A., Kron, S. J., Rios, G., Fink, G. R., & Serrano, R. (1995). Regulation of cation transport in Saccharomyces cerevisiae by the salt tolerance gene HAL3. *Molecular and Cellular Biology*, *15*(10), 5470-5481.

Ferreira A, Viljoen BC. (2003). Yeasts as adjunct starters in matured Cheddar cheese. Int J Food Microbiol 86: 131 – 140.

Fitzpatrick, D. A., Logue, M. E., Stajich, J. E., & Butler, G. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evolutionary Biology, 6(1), 99.

Flores, M., Durá, M. A., Marco, A., & Toldrá, F. (2004). Effect of Debaryomyces spp. on aroma formation and sensory quality of dry-fermented sausages. *Meat science*, *68*(3), 439-446.

Gadd, G. M., & Edwards, S. W. (1986). Heavy-metal-induced flavin production by Debaryomyces hansenii and possible connexions with iron metabolism. *Transactions of the British Mycological Society*, *87*(4), 533-542.

Gallardo, G., Ruiz-Moyano, S., Hernández, A., Benito, M. J., Córdoba, M. G., Pérez-Nevado, F., & Martín, A. (2014). Application of ISSR-PCR for rapid strain typing of Debaryomyces hansenii isolated from dry-cured Iberian ham. Food microbiology, 42, 205-211.

Gao, Q., Jin, K., Ying, S. H., Zhang, Y., Xiao, G., Shang, Y., ... & Peng, G. (2011). Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi Metarhizium anisopliae and M. acridum. PLoS Genet, 7(1), e1001264.

Gaxiola, R., De Larrinoa, I. F., Villalba, J. M., & Serrano, R. (1992). A novel and conserved salt-induced protein is an important determinant of salt tolerance in yeast. *The EMBO journal*, *11*(9), 3157.

Gırio, F. M., Amaro, C., Azinheira, H., Pelica, F., & Amaral-Collaço, M. T. (2000). Polyols production during single and mixed substrate fermentations in Debaryomyces hansenii. *Bioresource technology*, *71*(3), 245-251.

Gírio, F. M., Pelica, F., & Amaral-Collaço, M. T. (1996). Characterization of xylitol dehydrogenase from Debaryomyces hansenii. *Applied biochemistry and Biotechnology*, *56*(1), 79-87.

Gonzalez NA, Vázquez A, Ortiz Zuazaga HG, Sen A, Olvera HL, Peña de Ortiz S, Govind NS. 2009. Genome-wide expression profiling of the osmoadaptation response of Debaryomyces hansenii. Yeast. 26(2):111-124. In: PubMed [database on the internet]. PMID:19235772.

González-Hernández, J. C., Cárdenas-Monroy, C. A., & Pena, A. (2004). Sodium and potassium transport in the halophilic yeast Debaryomyces hansenii. *Yeast*, *21*(5), 403-412.

Gorjan, A., & Plemenitaš, A. (2006). Identification and characterization of ENA ATPases HwENA1 and HwENA2 from the halophilic black yeast Hortaea werneckii. *FEMS microbiology letters*, *265*(1), 41-50.

Groenewald, M., Daniel, H. M., Robert, V., Poot, G. A., & Smith, M. T. (2008). Polyphasic re-examination of Debaryomyces hansenii strains and reinstatement of D. hansenii, D. fabryi and D. subglobosus. Persoonia-Molecular Phylogeny and Evolution of Fungi, 21(1), 17-27.

Gumá-Cintrón, Y., Bandyopadhyay, A., Rosado, W., Shu-Hu, W., & Nadathur, G. S. (2015). Transcriptomic analysis of cobalt stress in the marine yeast Debaryomyces hansenii. *FEMS yeast research*, *15*(8).

Gunge, N., Fukuda, K., Morikawa, S., Murakami, K., Takeda, M., & Miwa, A. (1993). Osmophilic linear plasmids from the salt-tolerant yeast Debaryomyces hansenii. *Current genetics*, *23*(5-6), 443-449.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072-1075.

Hansen, K., Perry, B. A., Dranginis, A. W., & Pfister, D. H. (2013). A phylogeny of the highly diverse cup-fungus family Pyronemataceae (Pezizomycetes, Ascomycota) clarifies relationships and evolution of selected life history traits. Molecular Phylogenetics and Evolution, 67(2), 311-335.

Jacques, N., Zenouche, A., Gunde-Cimerman, N., & Casaregola, S. (2015). Increased diversity in the genus Debaryomyces from Arctic glacier samples. Antonie van Leeuwenhoek, 107(2), 487-501.

Jakobsen, M., & Narvhus, J. (1996). Yeasts and their possible beneficial and negative effects on the quality of dairy products. *International dairy journal*, *6*(8-9), 755-768.

James, T.Y., F. Kauff, C. Schoch, B. Matheny, V. Hofstetter, C.J. Cox, G. Celio, C. Guiedan, E. Fraker, J. Miadlikowska, T. Lumbsh, A. Rauhut, V. Reeb, A. Arnold, A. Amtoft, J.E. Stajich, K. Hosaka, G. Sung, D. Johnson, B. O'Rourke, M. Crockett, M. Binder, J.M. Curtis, J.C. Slot, Z. Wang, A.W. Wilson, A. Schueller, J.E. Longcore, K.O. Donnell, S. Mozley-Standridge, D. Porter, P.M. Letcher, M.J. Powell, J. W. Taylor, M.M. White, G.W. Griffith, D.R. Davies, R.A. Humber, J.B. Morton, J. Sugiyama, A.Y. Rossman, J.D. Rogers, D.H. Pfister, D. Hewitt, K. Hansen, S. Hambleton, R. A. Shoemaker, J. Kohlmeyer, B. Volkmann-Kohlmeyer, R.A. Spotts, M. Serdani, P.W. Crous, K.W. Hughes, K. Matsuura, E. Langer, G. Langer, W.A. Untereiner, R. Lucking, B. Budel, D.M. Geiser, D.M. Aptroot, P. Diederich, I. Schmitt, M. Schultz, R. Yahr, D.S. Hibbett, F. Lutzoni, D.J. Mclaughlin, J. W. Spatafora and R. Vilgalys. (2006). Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature 443, 818–822.

Jones, T., Federspiel, N. A., Chibana, H., Dungan, J., Kalman, S., Magee, B. B., ... & Davis, R. W. (2004). The diploid genome sequence of Candida albicans. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(19), 7329-7334.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772-780.

Kawaguchi, Y., Honda, H., Taniguchi-Morimura, J., & Iwasaki, S. (1989). The codon CUG is read as serine in an asporogenic yeast Candida cylindracea.

Kumar, S., Randhawa, A., Ganesan, K., Raghava, G. P. S., & Mondal, A. K. (2012). Draft genome sequence of salt-tolerant yeast Debaryomyces hansenii var. hansenii MTCC 234. *Eukaryotic cell*, *11*(7), 961-962.

Kuramae, E. E., Robert, V., Snel, B., Weiß, M., & Boekhout, T. (2006). Phylogenomics reveal a robust fungal tree of life. FEMS Yeast Research, 6(8), 1213-1220.

Kurtzman, C. P., & Robnett, C. J. (1998). Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. Antonie van Leeuwenhoek, 73(4), 331-371.

Kurtzman, C. P., & Robnett, C. J. (2003). Phylogenetic relationships among yeasts of the 'Saccharomyces complex'determined from multigene sequence analyses. FEMS yeast research, 3(4), 417-432.

Kurtzman, C. P., & Suzuki, M. (2010). Phylogenetic analysis of ascomycete yeasts that form coenzyme Q-9 and the proposal of the new genera Babjeviella, Meyerozyma, Millerozyma, Priceomyces, and Scheffersomyces. Mycoscience, 51(1), 2-14.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359.

Lopandic, K., Rentsendorj, U., Prillinger, H., & Sterflinger, K. (2013). Molecular characterization of the closely related Debaryomyces species: Proposition of D. vindobonensis sp. nov. from a municipal wastewater treatment plant. The Journal of general and applied microbiology, 59(1), 49-58.

Lorenz, R., & Molitoris, H. P. (1997). Cultivation of fungi under simulated deep sea conditions. *Mycological Research*, *101*(11), 1355-1365.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), pp-10.

Massey, S. E., Moura, G., Beltrão, P., Almeida, R., Garey, J. R., Tuite, M. F., & Santos, M. A. (2003). Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in Candida spp. Genome research, 13(4), 544-557.

Mattsson, R., Haemig, P. D., & Olsen, B. (1999). Feral pigeons as carriers of Cryptococcus laurentii, Cryptococcus uniguttulatus and Debaryomyces hansenii. *Medical Mycology*, *37*(5), 367-369.

Merdinger, E., & Devine, E. M. (1965). Lipids of Debaryomyces hansenii. *Journal of bacteriology*, *89*(6), 1488-1493.

Michán, C., Martínez, J. L., Alvarez, M. C., Turk, M., Sychrova, H., & Ramos, J. (2013). Salt and oxidative stress tolerance in Debaryomyces hansenii and Debaryomyces fabryi. FEMS yeast research, 13(2), 180-188.

Michán, C., Martínez, J. L., Alvarez, M. C., Turk, M., Sychrova, H., & Ramos, J. (2013). Salt and oxidative stress tolerance in Debaryomyces hansenii and Debaryomyces fabryi. *FEMS yeast research*, *13*(2), 180-188.

Mirarab, S., & Warnow, T. (2015). ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, *31*(12), i44-i52.

Mortensen, H. D., Gori, K., Jespersen, L., & Arneborg, N. (2005). Debaryomyces hansenii strains with different cell sizes and surface physicochemical properties adhere differently to a solid agarose surface. *FEMS microbiology letters*, *249*(1), 165-170.

Nakase, T., and M. Suzuki. (1985a). Taxonomic studies on Debaryomyces hansenii (Zopf) Lodder et Kreger-van Rij and related species. I. Chemotaxonomic investigations. J. Gen. Appl. Microbiol. 31, 49–69.

Nakase, T., and M. Suzuki. (1985b). Taxonomic studies on Debaryomyces hansenii (Zopf) Lodder et Kreger-van Rij and related species. II. Practical discrimination and nomenclature. J. Gen. Appl. Microbiol. 31, 71–86.

Navarrete, C., Siles, A., Martínez, J. L., Calero, F., & Ramos, J. (2009). Oxidative stress sensitivity in Debaryomyces hansenii. *FEMS yeast research*, *9*(4), 582-590.

Nguyen, H. V., Gaillardin, C., & Neuvéglise, C. (2009). Differentiation of Debaryomyces hansenii and Candida famata by rRNA gene intergenic spacer fingerprinting and reassessment of phylogenetic relationships among D. hansenii, C. famata, D. fabryi, C. flareri (= D. subglobosus) and D. prosopidis: description of D. vietnamensis sp. nov. closely related to D. nepalensis. FEMS yeast research, 9(4), 641-662.

Norkrans, B., & Kylin, A. (1969). Regulation of the potassium to sodium ratio and of the osmotic potential in relation to salt tolerance in yeasts. *Journal of Bacteriology*, *100*(2), 836-845.

Odds, F. C. (1988). *Candida and candidosis: a review and bibliography*. Bailliere Tindall.

Ohama, T., Suzuki, T., Mori, M., Osawa, S., Ueda, T., Watanabe, K., & Nakase, T. (1993). Non-universal decoding of the leucine codon CUG in several Candida species. *Nucleic acids research*, *21*(17), 4039-4045.

Parajó, J. C., Dominguez, H., & Domínguez, J. M. (1997). Improved xylitol production with Debaryomyces hansenii Y-7426 from raw or detoxified wood hydrolysates. *Enzyme and Microbial Technology*, *21*(1), 18-24.

Phaff, H. J., Vaughan-Martini, A., & Starmer, W. T. (1998). Debaryomyces prosopidis sp. nov., a yeast from exudates of mesquite trees. *International Journal of Systematic and Evolutionary Microbiology*, *48*(4), 1419-1424.

Prillinger, H., Molnár, O., Eliskases-Lechner, F., & Lopandic, K. (1999). Phenotypic and genotypic identification of yeasts from cheese. Antonie van Leeuwenhoek, 75(4), 267-283.

Prista, C., Almagro, A., Loureiro-Dias, M. C., & Ramos, J. (1997). Physiological basis for the high salt tolerance of Debaryomyces hansenii. Applied and Environmental Microbiology, 63(10), 4005-4009.

Prista, C., Loureiro-Dias, M. C., Montiel, V., García, R., & Ramos, J. (2005). Mechanisms underlying the halotolerant way of Debaryomyces hansenii. *FEMS yeast research*, *5*(8), 693-701.

Ratledge C, Tan K-H. 1990. In *Yeast Biotechnology and Biocatalysis*, Verachtert HJ, De Mot R (eds). Marcel Dekker: New York; 223 – 253.

Rios, G., Ferrando, A., & Serrano, R. (1997). Mechanisms of salt tolerance conferred by overexpression of the HAL1 gene in Saccharomyces cerevisiae. *Yeast*, *13*(6), 515-528.

Robbertse, B., Reeves, J. B., Schoch, C. L., & Spatafora, J. W. (2006). A phylogenomic analysis of the Ascomycota. Fungal genetics and biology, 43(10), 715-725.

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., ... & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, *61*(3), 539-542.

Rozpędowska, E., Piškur, J., & Wolfe, K. H. (2011). Genome sequences of Saccharomycotina: Resources and applications in phylogenomics. The Yeasts: A Taxonomic Study. London, UK: Elsevier, 5th edition, 145-157.

Ruiz, A., & Ariño, J. (2007). Function and regulation of the Saccharomyces cerevisiae ENA sodium ATPase system. *Eukaryotic cell*, *6*(12), 2175-2183.

Sánchez, N. S., Calahorra, M., González-Hernández, J. C., & Peña, A. (2006). Glycolytic sequence and respiration of Debaryomyces hansenii as compared to Saccharomyces cerevisiae. *Yeast*, *23*(5), 361-374.

Seda-Miro, J. M., Arroyo-Gonzalez, N., Perez-Matos, A., & Govind, N. S. (2007). Impairment of cobalt-induced riboflavin biosynthesis in a Debaryomyces hansenii mutant. *Canadian journal of microbiology*, *53*(11), 1272-1277.

Seda-Miró, Jasmine M. (2012). Phenotypic and Genetic Characterization of Debaryomyces hansenii Strains Exposed to Cobalt and Saline Stress (Doctoral dissertation, University of Puerto Rico at Mayaguez)

Seiler, H., & Busse, M. (1990). The yeasts of cheese brines. *International journal of food microbiology*, *11*(3-4), 289-303.

Shen, X. X., Zhou, X., Kominek, J., Kurtzman, C. P., Hittinger, C. T., & Rokas, A. (2016). Reconstructing the backbone of the Saccharomycotina yeast phylogeny using genome-scale data. G3: Genes| Genomes| Genetics, 6(12), 3927-3939.

Sherman, D. J., Martin, T., Nikolski, M., Cayla, C., Souciet, J. L., & Durrens, P. (2009). Génolevures: protein families and synteny among complete hemiascomycetous yeast proteomes and genomes. *Nucleic Acids Research*, *37*(Database issue), D550.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome research, 19(6), 1117-1123.

Stadler, J. A., & Schweyen, R. J. (2002). The yeast iron regulon is induced upon cobalt stress and crucial for cobalt tolerance. Journal of Biological Chemistry, 277(42), 39649-39654.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

Sugita, T., & Nakase, T. (1999). Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus Candida. *Systematic and applied microbiology*, *22*(1), 79-86.

Suzuki, M., Prasad, G. S., & Kurtzman, C. P. (2011). Debaryomyces Lodder & kreger-van rij (1952). The Yeasts: A Taxonomic Study. London, UK: Elsevier, 5th edition, 361.

Tafer, H., Sterflinger, K., & Lopandic, K. (2016). Draft genome of Debaryomyces fabryi CBS 789T, isolated from a human interdigital mycotic lesion. *Genome announcements*, *4*(1), e01580-15.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on mathematics in the life sciences, 17, 57-86.

Thomé-Ortiz, P. E., Pena, A., & Ramirez, J. (1998). Monovalent cation fluxes and physiological changes of Debaryomyces hansenii grown at high concentrations of KCl and NaCl. *Yeast*, *14*(15), 1355-1371.

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., ... & Roskies, R. (2014). XSEDE: accelerating scientific discovery. *Computing in Science & Engineering*, *16*(5), 62-74.

Ueda, T., Suzuki, T., Tokogawa, T., Nishikawa, K., & Watanabe, K. (1994). Unique structure of new serine tRNAs responsible for decoding leucine codon CUG in various Candida species and their putative ancestral tRNA genes. *Biochimie*, *76*(12), 1217-1222.

Wenning, M., Seiler, H., & Scherer, S. (2002). Fourier-transform infrared microspectroscopy, a novel and rapid tool for identification of yeasts. *Applied and environmental microbiology*, *68*(10), 4717-4721.

Wilmotte, A., Van De Peer, Y., Goris, A., Chapelle, S., De Baere, R., Nelissen, B., ... & De Wachter, R. (1993). Evolutionary relationships among higher fungi inferred from small ribosomal subunit RNA sequence analysis. Systematic and applied microbiology, 16(3), 436-444.

Wong, B., Kiehn, T. E., Edwards, F., Bernard, E. M., Marcove, R. C., De Harven, E., & Armstrong, D. (1982). Bone infection caused by debaryomyces hansenii in a normal host: a case report. *Journal of clinical microbiology*, *16*(3), 545-548.

Woolfit, M., E. Rozpe ḍowska, J. Piskur and K.H. Wolfe. (2007). Genome survey sequencing of the wine spoilage yeast Dekkera (Brettanomyces) bruxellensis. Eukaryot. Cell 6, 721–733.

Yanai, T., Tsunekawa, H., Okamura, K., & Okamoto, R. (1994). Manufacture of pyruvic acid with Debaryomyces. *Japanese Patent, JP 0,600,091*.

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics, 13(5), 329-342.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular evolution, 39(3), 306-314.

Yang, Z. (1996). Among-site rate variation and its impact on phylogenetic analyses. Trends in Ecology & Evolution, 11(9), 367-372.

APENDIX

Shell script for selecting the candidate genes for phylogenetic analysis.

```
#!/usr/bin/bash
#
#
#       Script to select the genes for phylogenetic studies using strain 767 (hansenii) as reference
#       using the nomenclature of the annotations provided by Maker.
#
#       It needs three python scripts:
#           rename_sequences.py will rename sequences for a simpler format
#           extract_sequences.py will extract sequences to a file
#           extract_final_sequences.py will extract sequences and append them together to a file
#
#
#       The following loop is going to search for the every gene of 767 to see if it is present in the other
#       strains.
#
#
for name in `cat yeast_list.txt`;do
            if [[ $name == 767 ]];then
                        cat $name.maker.proteins_functional_blast.fasta | grep ">" | cut -f 3 -d "=" | cut -f 1 -d " " |\
        sed 's/^/>/g' >$name.protein_names.txt
                        ./rename_sequences.py $name.maker.proteins_functional_blast.fasta $name.protein_names.txt \
            $name.sequences.fasta
        for x in `cat gene_names_767.txt`;do
          for y in `cat $name.sequences.fasta | grep ">" | grep -w "$x"`;do
                echo $x >>$name.767_share
          done
        done
        cat $name.767_share | sort -u >$name.767_share.txt
        cat $name.767_share.txt | sed 's/^/>/g' >$name.767_share.fasta
            else
        cat $name.maker.proteins_functional_blast.fasta | grep ">" | cut -f 1 -d "/" | \
             awk '{print $5}' | sed 's/^/>/g' >$name.protein_names.txt
            ./rename_sequences.py $name.maker.proteins_functional_blast.fasta $name.protein_names.txt \
         $name.sequences.fasta
        for x in `cat gene_names_767.txt`;do
        for y in `cat $name.sequences.fasta | grep ">" | grep -w "$x"`;do
                echo $x >>$name.767_share
          done
        done
        cat $name.767_share | sort -u >$name.767_share.txt
        cat $name.767_share.txt | sed 's/^/>/g' >$name.767_share.fasta
        fi
done
#
#
#       This loop is going to find the name of the genes of 767 that are share in all
#       species and save it in a new file
#
#
for x in `cat gene_names_767.txt`;do
        if [[ $x == `cat 1449.767_share.txt | grep -w "$x"` ]] && [[ $x == `cat 155.767_share.txt | grep -w "$x"`                    ]] && [[ $x
== `cat 398.767_share.txt | grep -w "$x"` ]] && [[ $x == `cat 7426.767_share.txt | grep –w       "$x"` ]] && [[ $x == `cat j26.767_share.txt |
grep -w "$x"` ]];then
                        echo $x >>genes_to_align_all_strains
    fi
done
#
#
#       Modify the file as a fasta file for using in python
#
cat genes_to_align_all_strains | sed 's/^/>/g' >genes_to_align_all_strains.txt
#
#
```

```
#       Loop to extract the shared genes in a new file and modify the fasta name with the specific strain name
#
for name in `cat yeast_list.txt`;do
    ./extract_sequences.py $name.sequences.fasta genes_to_align_all_strains.txt $name.sequences.to.align.fasta
    cat $name.sequences.to.align.fasta | sed "s/>/>$name./" | tr "." "_" >$name.align.fasta
done
#
#
#       Create a file for each gene and incorporate all the sequences of each strain. This will be used for alignments
#       Results will be store in the folder genes_to_align
#
mkdir genes_align
for x in `cat genes_to_align_all_strains`;do
    for name in `cat yeast_list.txt`;do
        if [[ `cat $name.align.fasta | grep ">" | cut -f 2 -d "_" | grep -wc "$x"` == 1 ]];then
                cat $name.align.fasta | grep ">" | grep -w "$name""_$x" >seq_extract.txt
            ./extract_final_sequences.py $name.align.fasta seq_extract.txt $x.gene.all.strains.fas
        else
            cat $name.align.fasta | grep ">" | grep -w "$name""_$x" >seq_extract.txt
            cat seq_extract.txt | while read line;do
                echo $line >seq_extract_uniq.txt
                ./extract_final_sequences.py $name.align.fasta seq_extract_uniq.txt \
            $x.gene.all.strains.fas
                break
            done
        fi
    done
done
mv *.fas genes_align/
#
#
#       Check if the files do not have duplicates
#
cd genes_align/
for x in `ls -1 *.fas`;do
        if [[ `cat $x | grep -c ">"` == 6 ]] && [[ `cat $x | grep ">" | cut -f 2 -d "_" | sort -u | wc -l` == 1 \                    ]];then
            echo $x >>no_duplicates.txt
        else
            echo $x >>check_duplicates.txt
        fi
done
#
#
#       Generate a new file with the 767 gene sequences(with not duplicates) shared in all other strains.
#
for x in `cat no_duplicates.txt`;do
        cat $x | grep ">" | grep "767_" >seq_extract.txt
        ../extract_final_sequences.py $x seq_extract.txt sequences_767_for_database.fasta
done
#
#
#       Blast of C.guilliermondii and 789 with 767 to get the shared genes for the final analysis.
#
#       Rename sequences
#
#       D.fabryi (789)
#
cp ../fabryi.proteins.fasta .
cat fabryi.proteins.fasta | grep ">" | cut -f 3 -d "=" | cut -f 1 -d "=" | cut -f 1 -d " " | sed 's/^/>/g' \ >fabryi.proteins_names.txt
../rename_sequences.py fabryi.proteins.fasta fabryi.proteins_names.txt 789.seq.fasta
cat 789.seq.fasta | sed 's/>/>789_/g' >789.sequences.fasta
#
#
#       C.guilliermondii
#
cp ../c.guilliermondii.maker.proteins_functional_blast.fasta  .
cat c.guilliermondii.maker.proteins_functional_blast.fasta  | grep ">" | cut -f 3 -d "=" | cut -f 1 -d "=" | \                    cut -f 1 -d " " | sed
's/^/>/g' >guilliermondii.proteins_names.txt
../rename_sequences.py c.guilliermondii.maker.proteins_functional_blast.fasta guilliermondii.proteins_names.txt \
```

```
guilliermondii.seq.fasta
cat guilliermondii.seq.fasta | sed 's/>/>guil_/g' >guilliermondii.sequences.fasta
#
#
#       Generate BLAST database
#
makeblastdb -in sequences_767_for_database.fasta -input_type fasta -dbtype prot
#
#
#       Blastp
#
for x in guilliermondii 789;do
        time blastp -db sequences_767_for_database.fasta -query $x.sequences.fasta -out $x.blastp -evalue .00001 \
        -outfmt 6 -num_alignments 1 -seg yes -soft_masking true -lcase_masking -max_hsps 1 -num_threads 15
done
#
#
#Extract the names of guilliermondii and 789, using 767 as reference.
#
cat no_duplicates.txt | cut -f 1 -d "." >767.names_extract.fasta
for x in `cat 767.names_extract.fasta`;do
        if [[ `cat guilliermondii.blastp | grep -w "767_""$x" | awk '{print $1}' | wc -l` -eq 1 ]] && \
        [[ `cat 789.blastp | grep -w "767_""$x" | awk '{print $1}' | wc -l` -eq 1 ]];then
        #guilliermondii
            cat guilliermondii.blastp | grep -w "767_""$x" | awk '{print $1}' | sed 's/^/>/g' >blast_extract.txt
            ../extract_final_sequences.py guilliermondii.sequences.fasta blast_extract.txt guilliermondii.fasta
            ../extract_sequences.py guilliermondii.sequences.fasta blast_extract.txt gene.sequence
            cat $x.gene.all.strains.fas gene.sequence >$x.gene.all.strains.fas2
        #789
            cat 789.blastp | grep -w "767_""$x" | awk '{print $1}' | sed 's/^/>/g' >blast_extract.txt
            ../extract_final_sequences.py 789.sequences.fasta blast_extract.txt 789.fasta
            ../extract_sequences.py 789.sequences.fasta blast_extract.txt gene.sequence2
            cat $x.gene.all.strains.fas2 gene.sequence2 >$x.gene.all.strains.fas3

        elif [[ `cat guilliermondii.blastp | grep -w "767_""$x" | awk '{print $1}' | wc -l` -gt 1 ]] \
         && [[ `cat 789.blastp | grep -w "767_""$x" | awk '{print $1}' | wc -l` -gt 1 ]];then
        #guilliermondii
            cat guilliermondii.blastp | grep -w "767_""$x" | awk '{print $1}' | sed 's/^/>/g' >blast_extract.txt
            cat blast_extract.txt | while read line;do
                        echo $line >blast_extract_uniq.txt
                        ../extract_final_sequences.py guilliermondii.sequences.fasta blast_extract_uniq.txt \
                         guilliermondii.fasta
                        ../extract_sequences.py guilliermondii.sequences.fasta blast_extract_uniq.txt gene.sequence
                        cat $x.gene.all.strains.fas gene.sequence >$x.gene.all.strains.fas2
                        break
            done
        #789
            cat 789.blastp | grep -w "767_""$x" | awk '{print $1}' | sed 's/^/>/g' >blast_extract.txt
            cat blast_extract.txt | while read line;do
                        echo $line >blast_extract_uniq.txt
                        ../extract_final_sequences.py 789.sequences.fasta blast_extract_uniq.txt 789.fasta
                        ../extract_sequences.py 789.sequences.fasta blast_extract_uniq.txt gene.sequence2
                        cat $x.gene.all.strains.fas2 gene.sequence2 >$x.gene.all.strains.fas3
                        break
            done

        elif [[ `cat guilliermondii.blastp | grep -w "767_""$x" | awk '{print $1}' | wc -l` -gt 1 ]] && \                                    [[ `cat
789.blastp | grep -w "767_""$x" | awk '{print $1}' | wc -l` -eq 1 ]];then
        #guilliermondii
            cat guilliermondii.blastp | grep -w "767_""$x" | awk '{print $1}' | sed 's/^/>/g' >blast_extract.txt
            cat blast_extract.txt | while read line;do
                        echo $line >blast_extract_uniq.txt
                        ../extract_final_sequences.py guilliermondii.sequences.fasta blast_extract_uniq.txt \
                        guilliermondii.fasta
                        ../extract_sequences.py guilliermondii.sequences.fasta blast_extract_uniq.txt gene.sequence
                        cat $x.gene.all.strains.fas gene.sequence >$x.gene.all.strains.fas2
                        break
            done
        #789
```

```
            cat 789.blastp | grep -w "767_""$x" | awk '{print $1}' | sed 's/^/>/g' >blast_extract.txt
            ../extract_final_sequences.py 789.sequences.fasta blast_extract.txt 789.fasta
            ../extract_sequences.py 789.sequences.fasta blast_extract.txt gene.sequence2
            cat $x.gene.all.strains.fas2 gene.sequence2 >$x.gene.all.strains.fas3

        elif [[ `cat guilliermondii.blastp | grep -w "767_""$x" | awk '{print $1}' | wc -l` -eq 1 ]] && \                    [[ `cat
789.blastp | grep -w "767_""$x" | awk '{print $1}' | wc -l` -gt 1 ]];then
        #guilliermondii
            cat guilliermondii.blastp | grep -w "767_""$x" | awk '{print $1}' | sed 's/^/>/g' >blast_extract.txt
            ../extract_final_sequences.py guilliermondii.sequences.fasta blast_extract.txt guilliermondii.fasta
            ../extract_sequences.py guilliermondii.sequences.fasta blast_extract.txt gene.sequence
            cat $x.gene.all.strains.fas gene.sequence >$x.gene.all.strains.fas2
        #789
            cat 789.blastp | grep -w "767_""$x" | awk '{print $1}' | sed 's/^/>/g' >blast_extract.txt
            cat blast_extract.txt | while read line;do
                    echo $line >blast_extract_uniq.txt
                    ../extract_final_sequences.py 789.sequences.fasta blast_extract_uniq.txt 789.fasta
                    ../extract_sequences.py 789.sequences.fasta blast_extract_uniq.txt gene.sequence2
                    cat $x.gene.all.strains.fas2 gene.sequence2 >$x.gene.all.strains.fas3
                    break
            done
        else
             echo $x >>genes_not_guil_789.txt
        fi
done
#
#
#       Extract now the sequences of j6, 234 and 17914 because they matched better with 789.
#
cp ../other_yeast.txt .
for name in `cat other_yeast.txt`;do
        cp ../$name.maker.proteins_functional_blast.fasta .
        cat $name.maker.proteins_functional_blast.fasta | grep ">" | cut -f 1 -d "/" | awk '{print $5}' | \                  sed
's/^/>/g' >$name.protein_names.txt
        ../rename_sequences.py $name.maker.proteins_functional_blast.fasta $name.protein_names.txt \
         $name.sequences.fasta
        cat 789.fasta | grep ">" | sed 's/>789_/>/g' | sort -u >sequences_extract_789.txt
        ../extract_final_sequences.py $name.sequences.fasta sequences_extract_789.txt $name.share.seq.789.fas
done
#
#       Modify fasta header
#
cat j6.share.seq.789.fas | sed 's/>/>j6_/g' >j6.share.seq.789.fast
cat 234.share.seq.789.fas | sed 's/>/>234_/g' >234.share.seq.789.fast
cat 17914.share.seq.789.fas | sed 's/>/>17914_/g' >17914.share.seq.789.fast
#
#
#       Loop to add the genes of j6, 234 and 17914 to a file for later concatenate with the other yeast genes.
#
cat sequences_extract_789.txt | tr -d ">" >sequences_789_extract_other
for x in `cat sequences_789_extract_other`;do
        if [[ `cat j6.share.seq.789.fast | grep ">" | grep -w "j6_""$x" | wc -l` -eq 1 ]] && \
        [[ `cat 234.share.seq.789.fast | grep ">" | grep -w "234_""$x" | wc -l` -eq 1 ]] \
        && [[ `cat 17914.share.seq.789.fast | grep ">" | grep -w "17914_""$x" | wc -l` -eq 1 ]];then
        #j6
            cat j6.share.seq.789.fast | grep ">" | grep -w "j6_""$x" >gene_j6
            ../extract_sequences.py j6.share.seq.789.fast gene_j6 gene_j6_seq
            cat gene_j6_seq >$x.gene.all.strains.fas4
        #234
            cat 234.share.seq.789.fast | grep ">" | grep -w "234_""$x" >gene_234
            ../extract_sequences.py 234.share.seq.789.fast gene_234 gene_234_seq
            cat $x.gene.all.strains.fas4 gene_234_seq >$x.gene.all.strains.fas5
        #17914
            cat 17914.share.seq.789.fast | grep ">" | grep -w "17914_""$x" >gene_17914
            ../extract_sequences.py 17914.share.seq.789.fast gene_17914 gene_17914_seq
            cat $x.gene.all.strains.fas5 gene_17914_seq >$x.gene.all.strains.fas6
        fi
done
#
```

```
#
#           Loop to concatenate the genes of j6, 234 and 179149 with the others.
#
cat 789.blastp | awk '{print $1,$2}' >789.blastp.extract
for x in `cat sequences_789_extract_other`;do
            if [[ `cat 789.blastp.extract | grep -w "789_""$x" | wc -l` == 1 ]];then
            #767 name
                a=$( cat 789.blastp.extract | grep -w "789_""$x" | awk '{print $2}' | cut -f 2 -d "_" )
            #789 name
                b=$( cat 789.blastp.extract | grep -w "789_""$x" | awk '{print $1}' | cut -f 2,3 -d "_" )
                cat $a.gene.all.strains.fas3 $b.gene.all.strains.fas6 >$a.genes.fasta
            fi
done
#
#
#           Check which file has all the reads
#
for x in `ls -1 *.genes.fasta`;do
   a=$( cat $x | grep ">" | sort -u | wc -l )
   if [ $a -eq 11 ];then
      echo $x >>final_genes_list
   fi
done
#
#           Create and move final genes to a new folder
#
mkdir final_genes
for x in `cat final_genes_list`;do
   cp $x final_genes/$x
 done
#
#End of script. Check files created and proceed to the alignment
#
```

Python scripts called by the previous script.

- rename_sequences.py

```
#!/usr/bin/python
#
#           rename_sequences.py will rename a fasta file
#
#           It needs (as arguments):
#               1-fasta file with sequences with the name to be changed
#               2-file with the replacing names
#               3-output filename
#
#
import sys

arg1=sys.argv[1]
arg2=sys.argv[2]
arg3=sys.argv[3]

fasta=open(arg1)
names=open(arg2)
newfasta=open(arg3,'w')


for line in fasta:
      if line.startswith('>'):
            newname=names.readline()
            newfasta.write(newname)
      else:
          newfasta.write(line)

fasta.close()
names.close()
```

```
newfasta.close()
```

- extract_sequences.py

```
#!/usr/bin/python
#
#       extract_sequences.py will extract a fasta sequence from a file
#
#       It needs (as arguments):
#           1-fasta file with sequences
#           2-file with the name of sequence to be extracted
#           3-output filename
#
#       If modified in the for loop, it can be use to filter fasta files
#
#
from Bio import SeqIO
import sys

arg1=sys.argv[1]
arg2=sys.argv[2]
arg3=sys.argv[3]

filename = arg1
reads_dict=SeqIO.to_dict(SeqIO.parse(arg2, "fasta"))
with open(arg3, "w") as output_file:
    n=0
        for record in SeqIO.parse(filename, "fasta"):
          if record.id in reads_dict:
                #n=n+1
                SeqIO.write(record, output_file, "fasta")
          else:
                        #SeqIO.write(record, output_file, "fasta")
                n=n+1

print(n)
```

- extract_final_sequences.py

```
#!/usr/bin/python
#
#       extract_final_sequences.py will extract fasta sequences and append them to a new file.
#
#       It needs (as arguments):
#           1-fasta file with sequences
#           2-file with the names to be extracted  and appended
#           3-output filename
#
#       If modified in the for loop, it can be use to filter fasta files
#
#
from Bio import SeqIO
import sys

arg1=sys.argv[1]
arg2=sys.argv[2]
arg3=sys.argv[3]

filename = arg1
reads_dict=SeqIO.to_dict(SeqIO.parse(arg2, "fasta"))
with open(arg3, "a") as output_file:
    n=0
        for record in SeqIO.parse(filename, "fasta"):
          if record.id in reads_dict:
                #n=n+1
                SeqIO.write(record, output_file, "fasta")
          else:
                        #SeqIO.write(record, output_file, "fasta")
                n=n+1
```

```
#print(n)
```